Comparison of prediction intervals for crack growth based on random effects models

Dissertation
by
Maryam Emdadi Fard

Supervisor:
Prof. Dr. Christine Müller

A thesis presented for the degree of
PhD of Statistics
Submitted to the Faculty of Statistics of the
TU Dortmund

technische universität
dortmund

Dortmund, January 2018

Abstract

Linear and nonlinear mixed effects models are applied extensively in the study of repeated measurements and longitudinal data. In this thesis, we propose two linear random effects models and a nonlinear random effects model based on the Paris-Erdogan equation for describing the crack growth data of Virkler et al. (1979). We describe how such models can be applied to achieve the future prediction and prediction interval of the time, when the crack attains a specific length. We propose eleven new methods for prediction interval by extending the methods of Swamy (1971), Rao (1975), Liski and Nummi (1996), Pinheiro and Bates (2000) and Stirnemann et al. (2011). We compare the methods and models by applying them on the crack propagation and simulated data.

Acknowledgment

First of all, I would like to thank my supervisor Prof. Dr. Christine Müller for her continuous supporting of my PhD project, for her patience and immense knowledge and for providing a very friendly working atmosphere in her research group. Her guidance helped me in my research and writing of this thesis. I could not have imaged having a better advisor for my PhD study.
I wish to thank all my colleagues for sharing their large statistical experiences and being always supportive and friendly, particularly Christoph Kustos, Dr. Lisa Denecke and Sebastian Szugat.
I also want to thank Professor B. M. Hillberry and Eric J. Tuegel for providing us the so called Virkler data.
I wish to thankfully acknowledge the financial support of DAAD, particularly in the award of a scholarship which provided the necessary financial support of this research.
Finally, I would like to thank my parents, my husband's parent, my darling husband, Amin, my lovely son, Amirali, and to my lovely daughter, Zeinab, for creating a happy life for me, and for their understanding.

Contents

List of Figures

List of Tables

List of Algorithms

Table 1: List of Symbols

Symbol	Description
I	Number of series ($i = 0, 1, \cdots, I$).
N_i	Number of observations in each series ($j = 1, \cdots, N_i$).
N_0	Number of observations in the observed part of the new series.
x_{ij}	The crack length.
x_0^F	The crack length of the future observation from the new series.
$\boldsymbol{x}_0^P := (x_{01}, \cdots, x_{0N_0})^\intercal$	The covariate vector of the observed part of the new series.
$\boldsymbol{x}_{ij} := (1, h_1(x_{ij}), \cdots, h_{q-1}(x_{ij}))$	A vector of functions of x_{ij} where $h_s : \mathbb{R} \to \mathbb{R}$; for $s = 1, \cdots, q-1$.
$\boldsymbol{x}_0^F := (1, h_1(x_0^F), \cdots, h_{q-1}(x_0^F))$	A vector of functions of x_0^F.
$\boldsymbol{X}_i := (\boldsymbol{x}_{i1}^\intercal, \cdots, \boldsymbol{x}_{iN_i}^\intercal)^\intercal$	The $(N_i \times q)$- covariate matrix of the ith old series.
$\boldsymbol{X}_0^P := (\boldsymbol{x}_{01}^\intercal, \cdots, \boldsymbol{x}_{0N_0}^\intercal)^\intercal$	The $(N_0 \times q)$-covariate matrix of the observed part of the new series.
$\boldsymbol{X}_0 := (\boldsymbol{X}_0^{P\intercal}, \boldsymbol{x}_0^{F\intercal})^\intercal$	The $((N_0 + 1) \times q)$-covariate matrix of the new series.
Y_{ij}	Time at the crack length x_{ij}.
y_{ij}	The realization of Y_{ij}.
$Y_i := (Y_{i1}, \cdots, Y_{iN_i})^\intercal$	The $(N_i \times 1)-$vector of observations from the ith old series ($i = 1, \cdots, I$).
$y_i := (y_{i1}, \cdots, y_{iN_i})^\intercal$	The realization of Y_i.
$Y_0^P := (Y_{01}, \cdots, Y_{0N_0})^\intercal$	The $(N_0 \times 1)$-vector of the observed part of the new series.
$y_0^P := (y_{01}, \cdots, y_{0N_0})^\intercal$	The realization of Y_0^P.
Y_0^F	The future observation of the new series.
y_0^F	The realization of Y_0^F.

$Y_0 := (Y_0^{P\intercal}, Y_0^F)^\intercal$	The $((N_0+1)\times 1)$-vector of observations of the new series.
$\boldsymbol{Y} := (Y_1^\intercal, \cdots, Y_I^\intercal)^\intercal$	The $((\sum_{i=1}^I N_i)\times 1)$-vector of observations from the old series.
$\boldsymbol{y} := (y_1^\intercal, \cdots, y_I^\intercal)^\intercal$	The realization of $\boldsymbol{Y}$.
$\boldsymbol{Y}_{obs} := (Y_1^\intercal, \cdots, Y_I^\intercal, Y_0^{P\intercal})^\intercal$	The $((\sum_{i=0}^I N_i)\times 1)$-vector of observations from the old and new series.
$\boldsymbol{y}_{obs} := (y_1^\intercal, \cdots, y_I^\intercal, y_0^{P\intercal})^\intercal$	The realization of $\boldsymbol{Y}_{obs}$.
$E_i := (E_{i1}, \cdots, E_{iN_i})^\intercal$	The $(N_i\times 1)$-vector of random errors of the ith old series .
$e_i := (e_{i1}, \cdots, e_{iN_i})^\intercal$	The realization of E_i.
$E_0^P := (E_{01}, \cdots, E_{0N_0})^\intercal$	The $(N_0\times 1)$-vector of random errors related to the observed part of the new series.
$e_0^P := (e_{01}, \cdots, e_{0N_0})^\intercal$	The realization of E_0^P.
E_0^F	Random error related to the unobserved (future) part of the new series.
e_0^F	The realization of E_0^F.
$E_0 := (E_0^{P\intercal}, E_0^F)^\intercal$	The $((N_0+1)\times 1)$-vector of random errors of the new series.
$e_0 := (e_0^{P\intercal}, e_0^F)^\intercal$	The realization of E_0.
$A_i := (A_{0i}, A_{1i}, \cdots, A_{(q-1)i})$	Random effect of the ith series, which is distributed as $A_i \overset{iid}{\sim} \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$, $i \in \{0, 1, \cdots, I\}$.
$a_i := (a_{0i}, a_{1i}, \cdots, a_{(q-1)i})$	The realization of A_i.
$\boldsymbol{\delta}_i := (\delta_{1i}, \cdots, \delta_{qi})^\intercal$	A vector of random variables which is distributed as $\boldsymbol{\delta}_i \overset{iid}{\sim} \mathbb{N}_q(\boldsymbol{0}, \Sigma)$.
q	The dimension of random effect.
$\boldsymbol{\alpha} \in \mathbb{R}^q$	The mean vector of random effects.
$\Sigma \in \mathbb{R}^{q\times q}$	The variance covariance matrix of random effects.
σ^2	The variance of random errors.
$\boldsymbol{\beta} := (\boldsymbol{\alpha}, \Sigma, \sigma^2)$	The parameter vector.

$g_i : \mathbb{R}^q \times \mathbb{R}^{N_i} \to \mathbb{R}^{N_i}$	The vector of nonlinear functions of the parameters related to the ith old series.
$g : \mathbb{R}^q \times \mathbb{R} \to \mathbb{R}$	The nonlinear function of the parameters.
$g_0^P : \mathbb{R}^q \times \mathbb{R}^{N_0} \to \mathbb{R}^{N_0}$	The vector of nonlinear functions of the parameters related to the observed part of the new series.
$g_0 : \mathbb{R}^q \times \mathbb{R}^{N_0+1} \to \mathbb{R}^{N_0+1}$	The vector of nonlinear functions of the parameters related to the new series.
$\boldsymbol{I}_{N_i}$	The $(N_i \times N_i)$-identity matrix.
$\boldsymbol{I}_q$	The $(q \times q)$-identity matrix.
$\mathbb{E}(X)$	Expectation of a random variable (vector) X .
$\mathbb{V}\text{ar}(X)$	Variance of random variable X.
$\mathbb{C}\text{ov}(X)$	Variance Covariance matrix of random vector X.
$\mathbb{C}\text{ov}(X, Y)$	Covariance of two random variables (vectors) X and Y.

1 Introduction

All materials show fatigue process after a certain time when subjected to cyclic loads. Fatigue leads to weakening of materials, which, if not fixed, could have catastrophic consequences. For example, the collapse of a bridge because of lack of tolerance of load or breaking of railway wheels axles because of tearing which both of them originate from fatigue. Such incidents can be avoided by promptly taking reinforcement measures on time and about the production of vehicles by replacing the new spare parts on time. However for costs reasons, such measures cannot occur too often. Therefore, an accurate prediction of fatigue time is of great importance.

Fatigue processes which lead to the fraction of materials can be described by cracks initiation and crack growth models. So studying of crack growth plays an important role in anticipating life time of products. For this reason, many experiments have been done on different materials like steel, concrete and etc. for evaluating their crack initiation, crack growth and crack propagation. A comprehensive study was conducted by Virkler et al. (1979), where the crack growth was observed in 68 specimens, with 164 measurements in each specimen. In other words, in the Virkler data set, we have observations $y_{i1}, \cdots, y_{iN_i}$ for the ith individual, where y_{ij} is the time until the crack length becomes $x_{ij}, \quad j = 1, \cdots, N_i = 164$ for all $i = 1, \cdots, 68$. For modelling the crack growth in the Virkler data set, simple random models can be derived from the Paris-Erdogan equation (Sobczyk and Spencer, 1992). One such model can be achieved by adding to the deterministic solutions of the Paris-Erdogan equation a random error to obtain a linear or nonlinear regression model. Another model can be achieved by adding a stochastic error to obtain a stochastic differential equation, which would be more desirable.

For crack propagation, Herman et al. (2016) propose a general Bayesian approach for stochastic versions of deterministic growth models. They consider six growth models in the form of a stochastic differential equation and a nonlinear regression model, however, they show that none of them is superior to the models based on the Paris-Erdogan equation. Therefore, we consider the following linear and nonlinear regression models which are based on the Paris-Erdogan equation:

1. Linear model:

$$y_{ij} = \alpha_0 + \alpha_1 \log(x_{ij}) + e_{ij},$$

2. Nonlinear model:

$$y_{ij} = \alpha_0 + \alpha_1 x_{ij}^{\alpha_2} + e_{ij},$$

3. Linearized model of the nonlinear model:

$$y_{ij} = \alpha_0 + \alpha_1 x_{ij} + \alpha_2 x_{ij} \log(x_{ij}) + e_{ij},$$

where e_{ij} is an additive error and α_0, α_1, α_2 are the unknown parameters.
As can be seen in Figure 1, in the Virkler data set, not only observations in each series differ from one another, but the series are also different. To address these variabilities, Hermann et al. (2016) propose a hierarchical model using a Bayesian approach.
In this research, a frequentist method is applied such that we apply random effects models, i.e., the parameters α_0, α_1, α_2 are replaced by the individual parameters a_{0i}, a_{1i}, a_{2i} which are the realizations of random variables, A_{0i}, A_{1i}, A_{2i}. In the next section, an introduction to estimation and prediction procedures in the random effects models is presented.

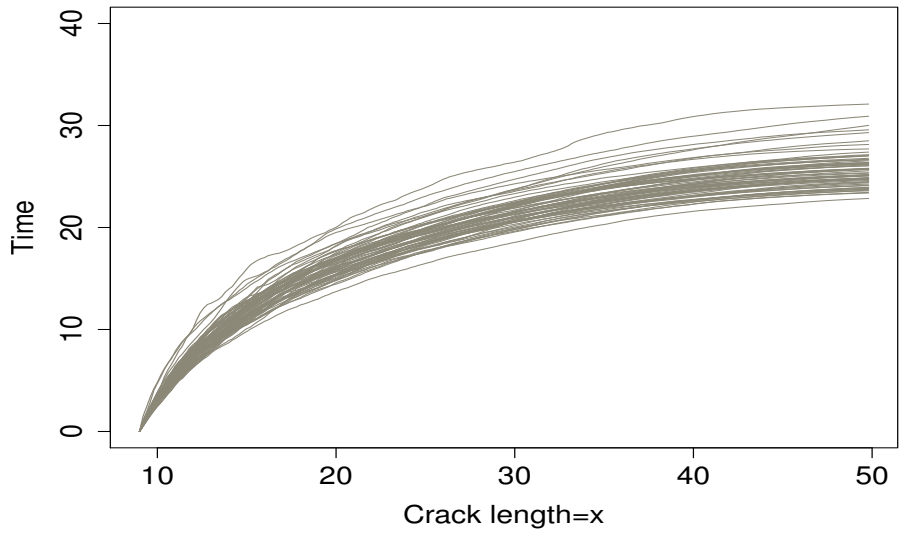

Figure 1: Virkler's data

1.1 Random effects models

Models with random and fixed effects are called mixed models. Hence we discuss here also the more general mixed effects models. Linear mixed models (LMMs) and nonlinear mixed models (NLMMs) have become emerged for analyzing data including repeated measurements. For estimation of parameters in the NLMMs and LMMs, several methods have been proposed (Davidian and Giltinan, 2003; Gumedze et al., 2011). In the NLMMs, maximum likelihood estimation of the unknown parameters is a considerable challenge as the likelihood of observations cannot be represented in a closed form. Therefore, some various approximations to the log-likelihood have been suggested such as the LME approximation (Lindstrom and Bates, 1990) and Laplace's approximation (Wolfinger, 1993; Vonesh, 1996; Wolfinger and Lin, 1997). Wang (2007) mentions that these approximations usually work well when the number of intraindividual measurements is large and the variability of random effects is small. By considering this weakness of the approximation-based methods, using more accurate methods such as the EM algorithm have been emerged. Particularly, the Monte Carlo Expectation Maximization (MCEM) algorithm (Wei and Tanner, 1990) is used such that the E step is approximated by the simulated samples from the conditional distribution of the random effects given the observed data. Walker (1996) suggests an MCEM algorithm for the exact maximum likelihood estimation in nonlinear random effects models such that the random effects are considered as the missing part of data. Kuhn and Lavielle (2005) propose a stochastic version of the EM algorithm (SAEM) by use of the stochastic approximations including samples which are obtained using the Markov chains. Jank (2006) proves that, unlike the MCEM, the SAEM algorithm is convergent with a fixed and commonly small simulation size. However, Wang (2007) mentions that in the NLMMs, the computational costs of independent samplers related to intractability of the target distribution can be significantly lower than corresponding value in the Markov chains. And this is despite the fact that the Markov chains can be applied to a wider variation of distributions than independent samples. Moreover, assessment of the Monte Carlo error for the MCEM algorithm based on the independent samples is straightforward whereas, it may be a complicated task for the MCEM algorithm based on the Markov chains. The Monte Carlo error should be evaluated after each EM iteration for checking whether it swamps the true EM step and if it happens the automatically increasing rule of the Monte Carlo sample size proposed by Booth and Hobert

(1999) is used. These advantages of the independent samplers over Markov chains motivate us to apply independent samplers in this research. In the most of research works (Walker, 1996; Kuhn and Lavielle, 2005; Wang, 2007), the considered missing parts of data in the EM algorithm are random effects and there are no missing observations or covariates in the data set. Wu (2004) proposes MCEM algorithms for the exact and approximate likelihood inferences, in the NLMMs, in the presence of missing covariates in the data and applies Gibbs sampling, importance sampling and rejection sampling for simulating required samples. In linear mixed models, Liski and Nummi (1996) propose an iterative EM algorithm by considering some missing observations in the data set as the missing part of the data.

In terms of prediction procedures in mixed models, Swamy (1971) proposes estimation and prediction procedures with prediction interval for the new observation from the new individual based on the information of the old individuals in random coefficient regression (RCR) models. Moreover, Liski and Nummi (1996) suggest a prediction procedure based on the EM algorithm and a prediction interval for the new observation from a partially observed individual in linear mixed models. In addition, Mathew et al. (2016) propose a prediction interval for the future observation by use of the generalized pivotal quantity (GPQ) percentages for the BLUP of the future observation in linear mixed effects models with a random effect. They suggest a GPQ for the BLUP of the future observation and then by generating Monte Carlo samples, they obtain 1000 GPQ values for the BLUP and finally present a prediction interval for the BLUP by use of the $\alpha/2$th and $(1-\alpha/2)$th percentiles of 1000 GPQ values. Furthermore, in the NLMMs, Hall and Clutter (2004) propose the linearization-based prediction procedure in the multilevel nonlinear mixed models and suggest an approximate prediction interval for the new observation from the partially observed individual. Moreover, Stirnemann et al. (2011) propose a simple prediction procedure given the past measurements of an individual by use of the Markov Chain Monte Carlo (MCMC) algorithm for simulating needed samples and suggest a prediction interval.
In the next section, the linear and nonlinear random effects models obtaining from the Paris-Erdogan equation and our new proposed methods applied in this thesis are presented.

1.2 Models and methods

We have the special case of the linear and nonlinear random effects models given as follows:

Linear random effects model for the old individuals:

$$Y_i := \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{iN_i} \end{pmatrix} = \boldsymbol{X}_i A_i + E_i = \begin{pmatrix} \boldsymbol{x}_{i1} \\ \vdots \\ \boldsymbol{x}_{iN_i} \end{pmatrix} A_i + \begin{pmatrix} E_{i1} \\ \vdots \\ E_{iN_i} \end{pmatrix},$$

with the realizations

$$y_i := \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iN_i} \end{pmatrix} = \boldsymbol{X}_i a_i + e_i = \begin{pmatrix} \boldsymbol{x}_{i1} \\ \vdots \\ \boldsymbol{x}_{iN_i} \end{pmatrix} a_i + \begin{pmatrix} e_{i1} \\ \vdots \\ e_{iN_i} \end{pmatrix},$$

where $y_i, e_i \in \mathbb{R}^{N_i}$, $\boldsymbol{X}_i \in \mathbb{R}^{N_i \times q}$, $a_i \in \mathbb{R}^q$, $\boldsymbol{x}_{ij} = (1, h_1(x_{ij}), \cdots, h_{q-1}(x_{ij}))$ (e.g. $h_1(x_{ij}) = x_{ij}, h_2(x_{ij}) = x_{ij}\log(x_{ij})$) with $h_s : \mathbb{R} \to \mathbb{R}, \quad s = 1, \cdots, q-1$, $i = 1, \cdots, I$.

Nonlinear random effects model for the old individuals:

$$Y_i := \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{iN_i} \end{pmatrix} = g_i(A_i, \boldsymbol{x}_i) + E_i = \begin{pmatrix} g(A_i, x_{i1}) \\ \vdots \\ g(A_i, x_{iN_i}) \end{pmatrix} + \begin{pmatrix} E_{i1} \\ \vdots \\ E_{iN_i} \end{pmatrix},$$

with the realizations

$$y_i := \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iN_i} \end{pmatrix} = g_i(a_i, \boldsymbol{x}_i) + e_i = \begin{pmatrix} g(a_i, x_{i1}) \\ \vdots \\ g(a_i, x_{iN_i}) \end{pmatrix} + \begin{pmatrix} e_{i1} \\ \vdots \\ e_{iN_i} \end{pmatrix},$$

where $y_i, e_i \in \mathbb{R}^{N_i}$, $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{iN_i})^\intercal \in \mathbb{R}^{N_i}$, $a_i \in \mathbb{R}^q$, $g : \mathbb{R}^q \times \mathbb{R} \to \mathbb{R}$, $g_i : \mathbb{R}^q \times \mathbb{R}^{N_i} \to \mathbb{R}^{N_i}$. The function g is nonlinear in a_i (e.g. $g(a_i, x_{ij}) = a_{0i} + a_{1i}x_{ij}^{a_{2i}}$) and $i = 1, \cdots, I$.

Linear and nonlinear random effects models for a new individual: For the new individual, we supposed that there are the N_0 past observations

and one future observation. Moreover, for the new individual, we have linear and nonlinear random effects models similar to the models for the old individuals, but we only need to partition vectors Y_0, E_0 and $g_0(A_0, \boldsymbol{x}_0)$ and matrix $\boldsymbol{X}_0$ into the past and future parts as follows:

$$Y_0 = (Y_0^{P\intercal}, Y_0^F)^\intercal, \quad E_0 = (E_0^{P\intercal}, E_0^F)^\intercal, \; g_0(A_0, \boldsymbol{x}_0) = (g_0^{P\intercal}(A_0, \boldsymbol{x}_0^P), g(A_0, x_0^F))^\intercal$$

and $\boldsymbol{X}_0 = (\boldsymbol{X}_0^{P\intercal}, \boldsymbol{x}_0^{F\intercal})^\intercal$, where $E_0^P = (E_{01}, \cdots, E_{0N_0})^\intercal$, $Y_0^P = (Y_{01}, \cdots, Y_{0N_0})^\intercal$ and $g_0^{P\intercal}(A_0, \boldsymbol{x}_0^P) = (g(A_0, x_{01}), \cdots, g(A_0, x_{0N_0}))^\intercal$ and indices P and F represent the past (observed) and future (unobserved) parts of the new individual, respectively. Let the realizations of Y_0^P, E_0 and E_0^P be defined as $y_0^P = (y_{01}, \cdots, y_{0N_0})^\intercal$, $e_0 = (e_0^{P\intercal}, e_0^F)^\intercal$ and $e_0^P = (e_{01}, \cdots, e_{0N_0})^\intercal$, respectively.
Finally, for both random effects models (i.e., linear and nonlinear models), we assume that $A_0, A_1, \cdots, A_I, E_0^P, E_0^F, E_1, \cdots, E_I$ are independent with $A_i \sim \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$, $E_i \sim \mathbb{N}_{N_i}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{N_i})$ for $i = 1, \cdots, I$ and $A_0 \sim \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$, $E_0 \sim \mathbb{N}_{N_0+1}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{N_0+1})$.

Let $\boldsymbol{Y} = (Y_1^\intercal, \cdots, Y_I^\intercal)^\intercal$ and $\boldsymbol{Y}_{obs} = (Y_1^\intercal, Y_I^\intercal, \cdots, Y_0^{P\intercal})^\intercal$ with the corresponding realizations $\boldsymbol{y} = (y_1^\intercal, \cdots, y_I^\intercal)^\intercal$ and $\boldsymbol{y}_{obs} = (y_1^\intercal, y_I^\intercal, \cdots, y_0^{P\intercal})^\intercal$. The aim of this thesis is to construct a prediction interval for a future observation Y_0^F at x_0^F of a new individual based on $\boldsymbol{Y}$ or $\boldsymbol{Y}_{obs}$, where observations $Y_{01}, \cdots, Y_{0N_0}$ at $x_{01}, \cdots, x_{0N_0}$ are given and x_0^F must be larger than x_{0N_0}.

In this thesis, eleven methods based on the modifications and extensions of the existing methods are presented and compared. It should be mentioned that some of these methods can only be used for the balanced data set, in which $N_i = N$, $\boldsymbol{X}_i = \boldsymbol{X}$ for $i = 1, \cdots, I$. In the following, we describe them.

1. **mSwamy (The modified method of Swamy for linear models in the balanced case (Swamy, 1971))**:
In the Swamy's prediction, the future observation Y_0^F at x_0^F is predicted as $\hat{y}_0^{F(old)} = \boldsymbol{x}_0^F \hat{\boldsymbol{\alpha}}_{GLS}$, where $\boldsymbol{x}_0^F = (1, h_1(x_0^F), \cdots, h_{q-1}(x_0^F))$ (e.g. $h_1(x_0^F) = x_0^F, h_2(x_0^F) = x_0^F \log(x_0^F)$) and $\hat{\boldsymbol{\alpha}}_{GLS}$ is the estimate of $\boldsymbol{\alpha}$ (i.e., the mean of random effects) which is obtained from the Swamy's estimation method based on the information of $\boldsymbol{Y}$ (the old individuals). A modification of the Swamy's prediction can be given by a convex combination of the Swamy's prediction

approach and the classical approach for linear models as follows:

$$\hat{y}_0^F = (1-w)\hat{y}_0^{F(old)} + w\hat{y}_0^{F(new)},$$

where w is chosen such that the variance of prediction error $(\hat{y}_0^F - y_0^F)$ is minimized, and $\hat{y}_0^{F(old)}$ is the Swamy's prediction, and $\hat{y}_0^{F(new)} = \boldsymbol{x}_0^F \hat{a}_{0,OLS}$ with $\hat{a}_{0,OLS} := (\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal} y_0^P$, where y_0^P is the realization of the observed part of the new individual (Y_0^P) and $\boldsymbol{X}_0^P = (\boldsymbol{x}_{01}^\intercal, \cdots, \boldsymbol{x}_{0N_0}^\intercal)^\intercal$ with $\boldsymbol{x}_{0j} = (1, h_1(x_{0j}), \cdots, h_{q-1}(x_{0j}))$ (e.g. $h_1(x_{01}) = x_{0j}, h_2(x_{0j}) = x_{0j}\log(x_{0j})$). In Lemma 4.2, we propose the $(1-\alpha)$-prediction interval for Y_0^F based on the modified method of Swamy.

2. **eRao (Extension of the method of Rao for linear models in the balanced case (Rao, 1975))**:
In the Rao's prediction, the prediction of Y_0^F at x_0^F is given by

$$\hat{y}_0^F = \boldsymbol{x}_0^F(\hat{C}\hat{a}_{0,OLS} + (\boldsymbol{I}_q - \hat{C})\hat{\boldsymbol{\alpha}}_{GLS}),$$

where $\boldsymbol{x}_0^F$, $\hat{a}_{0,OLS}$, $\hat{\boldsymbol{\alpha}}_{GLS}$ were defined as in the modified method of Swamy, $\hat{C} := \hat{\Sigma}(\hat{\Sigma} + \hat{\sigma}^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})^{-1}$, where $\hat{\Sigma}$ and $\hat{\sigma}^2$ are the estimates of Σ and σ^2 and are obtained from the Swamy's estimation method based on the information of $\boldsymbol{Y}$, and $\boldsymbol{X}_0^P$ is the covariate matrix related to the observed part of the new individual. Rao (1975) presents only the above prediction procedure and does not calculate prediction interval. We extend this approach by adding an approximate prediction interval for Y_0^F. In Corollary 4.1, we present this approximate prediction interval.

3. **mRao (The modified method of Rao for linear models in the balanced case (Rao, 1975))**:
In this approach, we consider a modification of the Rao's prediction such that $\boldsymbol{Y}_{obs}$ instead of $\boldsymbol{Y}$ is used for the estimation of parameters. Then a modified Rao's prediction is given by:

$$\hat{y}_0^F = \boldsymbol{x}_0^F(\hat{C}\hat{a}_{0,OLS} + (\boldsymbol{I}_q - \hat{C})\hat{\boldsymbol{\alpha}}_{obs})$$

where $\hat{C} := \hat{\Sigma}_{obs}[\hat{\Sigma}_{obs} + \hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}]^{-1}$, with $\hat{\boldsymbol{\alpha}}_{obs}$, $\hat{\Sigma}_{obs}$ and $\hat{\sigma}^2_{obs}$ which are obtained based on the information of $\boldsymbol{Y}_{obs}$. In Lemma 4.1 and Theorem 4.1, we calculate the distribution of prediction error and present an approximate prediction interval for Y_0^F based on the modified method of Rao, respectively.

4. **PBmSwamy (Combination of method of Pinheiro/Bates and the modified method of Swamy in the unbalanced case (Swamy, 1971; Pinheiro and Bates, 2000))**:
In this method, we linearize the nonlinear mixed model with the starting values, which are obtained from the estimation method of Pinheiro and Bates, and then use the linearized model to obtain the estimates and prediction intervals with the modified method of Swamy (see Section 4.2.2).

5. **ePB (Extension of the method of Pinheiro and Bates for nonlinear models in the unbalanced case (Pinheiro and Bates, 2000))**:
In the Pinheiro and Bates' method, the prediction of future observation (Y_0^F) at x_0^F is given by $\hat{y}_0^F = g(\hat{\boldsymbol{\alpha}}, x_0^F)$, where function g is a nonlinear and differentiable function of $\hat{\boldsymbol{\alpha}}$, and $\hat{\boldsymbol{\alpha}}$ is the estimate of $\boldsymbol{\alpha}$, which is obtained from the Pinheiro and Bates' estimation method based on the information of $\boldsymbol{Y}$ (the old individuals). We extend this approach by adding an approximate prediction interval for Y_0^F in Theorem 3.4.

6. **mPB (The modified method of Pinheiro and Bates for nonlinear models in the unbalanced case (Pinheiro and Bates, 2000))**:
In this approach, we present a modification of the Pinheiro and Bates' prediction such that the estimation of parameters are obtained based on the information of $\boldsymbol{Y}_{obs}$ (the old and new individuals) instead of $\boldsymbol{Y}$. Based on this approach, we propose an approximate prediction interval for Y_0^F in Theorem 4.2.

7. **PBLiski (Combination of methods of Pinheiro/Bates and Liski/Nummi in the unbalanced case (Liski and Nummi, 1996; Pinheiro and Bates, 2000))**:
Liski and Nummi (1996) propose an EM algorithm to calculate the maximum likelihood estimator for the parameters in a linear mixed model and derive a prediction interval for Y_0^F. As in the Method 4, we propose to start with the estimates of Pinheiro and Bates, linearize the nonlinear model with these estimates and then use the linearized model to obtain the estimates and prediction intervals with the method of Liski and Nummi (see Section 4.2.1). In Lemma 4.3, the approximate prediction interval in a linear mixed model proposed by Liski and Nummi (1996) is presented.

8. **PBSt (Combination of methods of Pinheiro/Bates and Stirnemann et al. in the unbalanced case (Stirnemann et al., 2011; Pinheiro and Bates, 2000))**:
Let $\hat{\boldsymbol{\beta}}$ be the ML estimation of $\boldsymbol{\beta} := (\boldsymbol{\alpha}, \Sigma, \sigma^2)$ based on the information of $\boldsymbol{Y}$, which is obtained from the estimation method of Pinheiro and Bates. In the method of Stirnemann et al., the future observation is predicted as:

$$\hat{y}_0^F := \mathbb{E}(\widehat{Y_0^F | y_0^P}, \hat{\boldsymbol{\beta}}) \approx \frac{1}{M} \sum_{m=1}^{M} g(\tilde{a}_0^{(m)}, x_0^F),$$

where M is the Monte Carlo sample size, g is a nonlinear and differentiable function of $\tilde{a}_0^{(m)}$ and $\tilde{a}_0^{(m)}$ for $m \in \{1, \cdots, M\}$ is the random sample which is obtained using the MCMC algorithm from $f(a_0|y_0^P, \hat{\boldsymbol{\beta}})$, where a_0 is the realization of A_0, the random effect related to the new individual (see Lemma 4.5). Stirnemann et al. (2011) propose the $(1-\alpha)$-prediction interval for Y_0^F, which is given by

$$PI = [q_{\alpha/2}(\hat{\boldsymbol{\beta}}, y_0^P), q_{1-\alpha/2}(\hat{\boldsymbol{\beta}}, y_0^P)],$$

where $q_{\alpha/2}(\hat{\boldsymbol{\beta}}, y_0^P)$ and $q_{1-\alpha/2}(\hat{\boldsymbol{\beta}}, y_0^P)$ are respectively the $(\alpha/2)-$ and $(1-\alpha/2)-$ quantiles of $g(\tilde{a}_0^{(1)}, x_0^F), \cdots, g(\tilde{a}_0^{(M)}, x_0^F)$. For simulating samples using the MCMC algorithm from $f(a_0|y_0^P, \boldsymbol{\beta})$, the special R package was used (see Section 4.2.5.1).

9. **StConf (Method of Stirnemann et al. combined with confidence sets in the unbalanced case (Stirnemann et al., 2011))**:
In this approach, we use the same estimation and prediction procedures which have been used in Method 8 and consider a prediction interval based on the $(1-\alpha)$-confidence ellipsoid of parameters, which results in a $(1-\alpha)^2$-prediction interval (Müller et al., 2016), as follows:

$$\cup_{\boldsymbol{\beta} \in \hat{\Theta}(\boldsymbol{y})} [q_{\alpha/2}(\boldsymbol{\beta}, y_0^P), q_{1-\alpha/2}(\boldsymbol{\beta}, y_0^P)] \subseteq [\min_{\boldsymbol{\beta} \in \hat{\Theta}(\boldsymbol{y})} q_{\alpha/2}(\boldsymbol{\beta}, y_0^P), \max_{\boldsymbol{\beta} \in \hat{\Theta}(\boldsymbol{y})} q_{1-\alpha/2}(\boldsymbol{\beta}, y_0^P)],$$

where $q_{\alpha/2}(\boldsymbol{\beta}, y_0^P)$ and $q_{1-\alpha/2}(\boldsymbol{\beta}, y_0^P)$ are respectively the $(\frac{\alpha}{2})$-and $(1-\frac{\alpha}{2})$-quantiles of $g(\tilde{a}_0^{(1)}, x_0^F), \cdots, g(\tilde{a}_0^{(M)}, x_0^F)$ for each $\boldsymbol{\beta}$ in the confidence ellipsoid $\hat{\Theta}(\boldsymbol{y})$ and $\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)}$ is a random sample from $f(a_0|y_0^P, \boldsymbol{\beta})$ (see Section 4.2.5.2). Similar to Method 8, an MCMC algorithm is used for simulating

samples from $f(A_0|y_0^P, \boldsymbol{\beta})$. The confidence ellipsoid for $\boldsymbol{\beta}$ is obtained from the asymptotic normality of the estimates obtaining from the Pinheiro and Bates' method (see Section 3.2.1.1).

10. **WmSt (Combination of an extended method of Walker and a modified method of Stirnemann et al. in the unbalanced case (Walker, 1996; Stirnemann et al., 2011))**:
Walker (1996) proposes an MCEM algorithm for estimating the parameters of a nonlinear random effects model in which $\boldsymbol{Z} := (\boldsymbol{Y}_{obs}, \boldsymbol{A})$ is the complete data, where $\boldsymbol{A} := (A_1, \cdots, A_I, A_0)$, which is a vector of all random effects, is the missing part and $\boldsymbol{Y}_{obs}$ is a vector of the observed parts of all individuals. We extend the Walker's MCEM algorithm such that $\boldsymbol{Z} := (\boldsymbol{Y}_{obs}, \boldsymbol{A}, Y_0^F)$ is the complete data and $\boldsymbol{A}$ and Y_0^F are the missing parts, where Y_0^F is the unknown observation from the new individual that is partially observed. For simulating samples, which is needed in the Monte Carlo approximation, we apply the Walker's method for the completely observed individuals and propose to apply the Gibbs sampling along with the Sampling Importance Resampling (SIR) method for the partially observed individual (Koch, 2007; Bishop, 2006) (see Algorithm 1). The pseudocode for our proposed MCEM algorithm is presented in Algorithm 2.
For prediction of the future observation, we suggest a modified method of Stirnemann et al. in which the ratio of two Monte Carlo approximations, proposed by Walker (1996), is used for calculating the conditional expectations and predictions. It should be mentioned that, as the simulated samples from the prediction method of Stirnemann et al. are obtained from an MCMC algorithm, so they have dependent samples in this case. However in the modified version of this approach, the samples are independent (see Section 4.2.6.2). A modified prediction of Strinemann et al. is given by:

$$\hat{y}_0^F := \mathbb{E}(\widehat{Y_0^F|y_0^P}, \hat{\boldsymbol{\beta}}) \approx \sum_{m=1}^{M} g(\tilde{a}_0^{(m)}, x_0^F) \frac{f(y_0^P|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})}{\sum_{m=1}^{M} f(y_0^P|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})},$$

where $\hat{\boldsymbol{\beta}}$ is the ML estimation of $\boldsymbol{\beta} := (\boldsymbol{\alpha}, \Sigma, \sigma^2)$ based on the information of $\boldsymbol{Y}_{obs}$, which is obtained from the extension of Walker's estimation method, and $(\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)})$ is the random sample from $f(a_0|\hat{\boldsymbol{\beta}})$ $((A_0|\hat{\boldsymbol{\beta}}) \sim \mathbb{N}_q(\hat{\boldsymbol{\alpha}}, \hat{\Sigma}))$. For prediction interval, we propose to use the following $(1-\alpha)$-prediction interval:

$$PI = [q_{\alpha/2}(\hat{\boldsymbol{\beta}}), q_{1-\alpha/2}(\hat{\boldsymbol{\beta}})],$$

where $q_{\alpha/2}(\hat{\boldsymbol{\beta}})$ and $q_{1-\alpha/2}(\hat{\boldsymbol{\beta}})$ are respectively the $(\alpha/2)-$ and $(1-\alpha/2)-$ quantiles of $g(\tilde{a}_0^{(1)}, x_0^F), \cdots, g(\tilde{a}_0^{(M)}, x_0^F)$, and $(\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)})$ is a random sample from $\mathbb{N}_q(\hat{\boldsymbol{\alpha}}, \hat{\Sigma})$ (see Section 4.2.6.3).

11. **mStConf (The modified method of Stirnemann et al. combined with confidence sets (Stirnemann et al., 2011))**:
In this method, we use the same estimation and prediction procedures which have been described in Method 10. But for prediction interval, based on the results in Müller et al. (2016), we propose the $(1-2\alpha)$-prediction interval by use of the $(1-\alpha)$-confidence set of parameters $\hat{\Theta}(\boldsymbol{y}_{obs})$, which is given by

$$\cup_{\boldsymbol{\beta}\in\hat{\Theta}(\boldsymbol{y}_{obs})}[q_{\alpha/2}(\boldsymbol{\beta}), q_{1-\alpha/2}(\boldsymbol{\beta})] \subseteq [\min_{\boldsymbol{\beta}\in\hat{\Theta}(\boldsymbol{y}_{obs})} q_{\alpha/2}(\boldsymbol{\beta}), \max_{\boldsymbol{\beta}\in\hat{\Theta}(\boldsymbol{y}_{obs})} q_{1-\alpha/2}(\boldsymbol{\beta})],$$

where $q_{\alpha/2}(\boldsymbol{\beta})$ and $q_{1-\alpha/2}(\boldsymbol{\beta})$ are respectively the $(\frac{\alpha}{2})$-and $(1-\frac{\alpha}{2})$-quantile of $g(\tilde{a}_0^{(1)}, x_0^F), \cdots, g(\tilde{a}_0^{(M)}, x_0^F)$ for each $\boldsymbol{\beta}$ in the confidence ellipsoid $\hat{\Theta}(\boldsymbol{y}_{obs})$, and $(\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)})$ is a random sample from $f(a_0|\boldsymbol{\beta})$ $((A_0|\boldsymbol{\beta}) \sim \mathbb{N}_q(\boldsymbol{\alpha}, \boldsymbol{\Sigma}))$ for each $(\boldsymbol{\alpha}, \Sigma)$ in $\hat{\Theta}(\boldsymbol{y}_{obs})$. The related confidence ellipsoid is obtained using Theorem 4.5 which has been proposed by Nie (2007) (see Section 4.2.6.4).

1.3 Thesis overview

This thesis is organized in 8 Chapters. In Chapter 2, the existing estimation and prediction methods in linear and nonlinear regression models suggested by Seber (1989) and Bates (1988) are discussed. In this chapter, for the estimation and prediction, only the information from the new series is used. Chapter 3 describes the existing estimation and prediction methods for linear and nonlinear random effects models suggested by Swamy (1971) and Pinheiro and Bates (2000). In this chapter, estimation of the parameters and prediction of the random effects and missing observations are obtained by use of the information of the old series. It also contains our proposed approximate prediction interval for Y_0^F based on the Pinheiro and Bates' prediction, which is presented in Theorem 3.4.
Chapter 4 introduces some of the existing prediction methods in linear and nonlinear random effects models. These methods are due to Rao (1975), Liski and Nummi (1996), Hall and Clutter (2004) and Stirnemann et al. (2011). This chapter also contains the modifications and extensions of some of the existing prediction methods (extension of the method of Rao, the modified

method of Rao, the modified method of Swamy, combination of method of Pinheiro/Bates and the modified method of Swamy, combination of methods of Pinheiro/Bates and Liski/Nummi, extension of the method of Pinheiro and Bates, the modified method of Pinheiro and Bates, combination of methods of Pinheiro/Bates and Stirnemann et al., method of Stirnemann et al. combined with confidence sets, combination of an extended method of Walker and a modified method of Stirnemann et al., and the modified method of Stirnemann et al. combined with confidence sets). The most important part of this chapter is Section 4.2. In this chapter, we propose some new lemmas (Lemma 4.1 and 4.2) and theorems (Theorem 4.1 and 4.2) for calculating distribution of the prediction error and getting the approximate prediction interval for Y_0^F.

The difference between Chapters 2, 3 and 4 is due to the information used in the estimation and prediction procedures. In Chapter 2, the estimation of unknown parameters and prediction of the future observation are obtained only based on the information of the new series (partially observed series), whereas in Chapter 3, in all the estimation and prediction procedures we use the information of the old series. However in Chapter 4, the information of the old and new series is used, which is our main aim in this thesis.

Chapter 5 contains numerical results from a simulation study and an application to a real data. In Chapter 6, the results are summarized and for the future work an outlook is also given. Chapters 7 and 8 include Appendix and a list of references.

2 Prediction using only observations of the same series

2.1 Prediction of the future observation in linear regression models

2.1.1 Estimation method of Seber

Consider the following linear regression model for the partially observed individual (new individual):

$$Y_0^P = \boldsymbol{X}_0^P \boldsymbol{\alpha} + E_0^P \tag{2.1}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^q$ and

$$Y_0^P = \begin{pmatrix} Y_{01} \\ . \\ . \\ . \\ Y_{0N_0} \end{pmatrix}, \quad \boldsymbol{X}_0^P = \begin{pmatrix} \boldsymbol{x}_{01} \\ \vdots \\ \boldsymbol{x}_{0N_0} \end{pmatrix}, \quad E_0^P = \begin{pmatrix} E_{01} \\ . \\ . \\ . \\ E_{0N_0} \end{pmatrix},$$

with $\boldsymbol{x}_{0j} = (1, h_1(x_{0j}), \cdots, h_{q-1}(x_{0j})), \quad j = 1, \cdots, N_0$. Y_0^P is an $N_0 \times 1$ vector of the observations, $\boldsymbol{X}_0^P$ is an $N_0 \times q$ covariate matrix, $\boldsymbol{\alpha}$ is a $q \times 1$ vector of the regression coefficients, and E_0^P is an $N_0 \times 1$ vector of the random errors which is distributed as $E_0^P \sim \mathbb{N}_{N_0}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{N_0})$.

Theorem 2.1. *Estimation of $\boldsymbol{\alpha}$ and σ^2 [Seber (1989)]:*
Consider the linear model (2.1). The unbiased estimations of $\boldsymbol{\alpha}$ and σ^2 are given by

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{X}_0^{P\intercal} Y_0^P \tag{2.2}$$

$$\hat{\sigma^2} = \frac{(Y_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}})^\intercal (Y_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}})}{N_0 - q}. \tag{2.3}$$

Moreover, it is proved that

$$\hat{\boldsymbol{\alpha}} \sim \mathbb{N}_q(\boldsymbol{\alpha}, \sigma^2 (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}) \tag{2.4}$$

By use of the *lm* function in the stats package in R, these estimates can be obtained (R Development Core Team, 2016).

2.1.2 Prediction method of Seber

In regression models, we are interested in predicting the future observation Y_0^F which is modeled as

$$Y_0^F = \boldsymbol{x}_0^F \boldsymbol{\alpha} + E_0^F, \tag{2.5}$$

where $\boldsymbol{x}_0^F = (1, h_1(x_0^F) \cdots, h_{q-1}(x_0^F))$ and $E_0^F \sim \mathbb{N}(0, \sigma^2)$ such that E_0^P and E_0^F are independent. Prediction of the future observation is given by

$$\hat{y}_0^F = \boldsymbol{x}_0^F \hat{\boldsymbol{\alpha}},$$

where $\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{X}_0^{P\intercal} y_0^P$ and y_0^P is the realization of Y_0^P.
By use of the *predict* function in the stats package in R, prediction and prediction interval for the future observation is obtained.

Theorem 2.2. *[Seber (1989)] :*
Let $U \sim \mathbb{N}_n(\boldsymbol{\mu}, \sigma^2 I_n)$ be an n-vector and B be an $n \times n$ symmetric matrix. Then the ratio $U'BU/\sigma^2$ will have a $\chi_r^2(\lambda)$ distribution with $\lambda = \boldsymbol{\mu}' B \boldsymbol{\mu}/\sigma^2$ if B is idempotent with rank(B)$= r$.

Theorem 2.3. *Prediction interval for Y_0^F [Seber (1989)]*
Consider the linear regression model (2.1) for the partially observed individual and the linear model (2.5) for the future observation from the new individual. Prediction of the future observation Y_0^F is given by

$$\hat{Y}_0^F = \boldsymbol{x}_0^F \hat{\boldsymbol{\alpha}},$$

where $\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{X}_0^{P\intercal} Y_0^P$. The prediction error $(\hat{Y}_0^F - Y_0^F)$ has normal distribution as follows:

$$\hat{Y}_0^F - Y_0^F \sim \mathbb{N}(0, \sigma^2[1 + \boldsymbol{x}_0^F (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{x}_0^{F\intercal}]), \tag{2.6}$$

and then the $(1-\alpha)$-prediction interval for Y_0^F can be given by

$$\mathcal{PI} = (\hat{y}_0^F \pm t_{N_0-q}(\alpha/2) \sqrt{\hat{\sigma}^2[1 + \boldsymbol{x}_0^F (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{x}_0^{F\intercal}]}), \tag{2.7}$$

where $\hat{y}_0^F$ is the realization of $\hat{Y}_0^F$.

Proof. The proof can be found in the appendix. □

2.2 Prediction of the future observation in nonlinear regression models

2.2.1 Gauss-Newton estimation method

Consider the following nonlinear regression model for the partially observed individual as:

$$Y_0^P = g_0^P(\boldsymbol{\alpha}, \boldsymbol{x}_0^P) + E_0^P, \tag{2.8}$$

where $g_0^P(\boldsymbol{\alpha}, \boldsymbol{x}_0^P) : \mathbb{R}^q \times \mathbb{R}^{N_0} \to \mathbb{R}^{N_0}$ is a vector of the nonlinear functions of $\boldsymbol{\alpha}$, $\boldsymbol{\alpha} \in \mathbb{R}^q$ and

$$Y_0^P = \begin{pmatrix} Y_{01} \\ . \\ . \\ . \\ Y_{0N_0} \end{pmatrix}, \quad g_0^P(\boldsymbol{\alpha}, \boldsymbol{x}_0^P) = \begin{pmatrix} g(\boldsymbol{\alpha}, x_{01}) \\ \vdots \\ g(\boldsymbol{\alpha}, x_{0N_0}) \end{pmatrix}, \quad E_0^P = \begin{pmatrix} E_{01} \\ . \\ . \\ . \\ E_{0N_0} \end{pmatrix}.$$

Y_0^P is an $N_0 \times 1$ vector of the observations, $\boldsymbol{x}_0^P \in \mathbb{R}^{N_0}$ is a regressor variables vector, $\boldsymbol{\alpha}$ is a $q \times 1$ vector of the regression coefficients, and E_0^P is an $N_0 \times 1$ vector of the random errors which is distributed as $E_0^P \sim \mathbb{N}_{N_0}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{N_0})$.

The Gauss-Newton method [Bates (1988)]: Gauss proposes using a linear approximation to the nonlinear function and iteratively improving the starting value $\boldsymbol{\alpha}^{(0)}$ for $\boldsymbol{\alpha}$. At the $(k+1)$th iteration, the first order Taylor expansion of $g_0^P(\boldsymbol{\alpha}, \boldsymbol{x}_0^P)$ about $\boldsymbol{\alpha}^{(k)}$ is given by

$$g_0^P(\boldsymbol{\alpha}, \boldsymbol{x}_0^P) \approx g_0^P(\boldsymbol{\alpha}^{(k)}, \boldsymbol{x}_0^P) + \tilde{\boldsymbol{X}}_0^{P(k+1)}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(k)}), \tag{2.9}$$

$$\tilde{\boldsymbol{X}}_0^{P(k+1)} = \frac{\partial g_0^P(\boldsymbol{a}, \boldsymbol{x}_0^P)}{\partial \boldsymbol{a}}|_{\boldsymbol{a}=\boldsymbol{\alpha}^{(k)}} \in \mathbb{R}^{N_0 \times q}.$$

This is equivalent to approximate the residuals at the $(k+1)$th iteration as follows:

$$\begin{aligned} E_0^P = Y_0^P - g_0^P(\boldsymbol{\alpha}, \boldsymbol{x}_0^P) &\approx Y_0^P - [g_0^P(\boldsymbol{\alpha}^{(k)}, \boldsymbol{x}_0^P) + \tilde{\boldsymbol{X}}_0^{P(k+1)}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(k)})] \\ &= Z_0^{P(k+1)} - \tilde{\boldsymbol{X}}_0^{P(k+1)}\boldsymbol{\delta}, \end{aligned} \tag{2.10}$$

where $Z_0^{P(k+1)} = Y_0^P - g_0^P(\boldsymbol{\alpha}^{(k)}, \boldsymbol{x}_0^P)$ and $\boldsymbol{\delta} = \boldsymbol{\alpha} - \boldsymbol{\alpha}^{(k)}$. In fact, at the $(k+1)$th iteration we have approximately a linear regression model as

$$Z_0^{P(k+1)} \approx \tilde{\boldsymbol{X}}_0^{P(k+1)}\boldsymbol{\delta} + E_0^P.$$

Hence, the least square estimate of $\boldsymbol{\delta}$ at the $(k+1)$ th iteration is given by

$$\boldsymbol{\delta}^{(k+1)} = (\tilde{\boldsymbol{X}}_0^{P(k+1)^\intercal} \tilde{\boldsymbol{X}}_0^{P(k+1)})^{-1} \tilde{\boldsymbol{X}}_0^{P(k+1)^\intercal} Z_0^{P(k+1)}.$$

Finally, the update estimate of $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\alpha}^{(k)} + \boldsymbol{\delta}^{(k+1)}.$$

This process is repeated until no useful changes appear in the elements of the parameter vector.
Consider $\hat{\boldsymbol{\alpha}}$ be the estimate from the final iteration of the Gauss-Newton method. Then the estimate of σ^2 is obtained by

$$\hat{\sigma}^2 = \frac{(Y_0^P - g_0^P(\hat{\boldsymbol{\alpha}}, \boldsymbol{x}_0^P))^\intercal (Y_0^P - g_0^P(\hat{\boldsymbol{\alpha}}, \boldsymbol{x}_0^P))}{N_0 - q}. \tag{2.11}$$

These estimates can be obtained from the *nls* function in the stats package in R (R Development Core Team, 2016).

Asymptotic normality of $\hat{\boldsymbol{\alpha}}$:
Jennrich (1969) proves that

$$\sqrt{N_0}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{N_0 \to \infty} \mathbb{N}_q(\mathbf{0}, \sigma^2 \boldsymbol{I}_{\boldsymbol{\alpha}}^{-1}), \tag{2.12}$$

where

$$\boldsymbol{I}_{\boldsymbol{\alpha}} := \int \dot{g}(\boldsymbol{\alpha}, x) \dot{g}(\boldsymbol{\alpha}, x)^\intercal \Delta(dx),$$

with $\dot{g}(\boldsymbol{\alpha}, x) := \frac{\partial g(\boldsymbol{a}, x)}{\partial \boldsymbol{a}}|_{\boldsymbol{a}=\boldsymbol{\alpha}}$ and Δ is a design measure.
From (2.12), we approximately have

$$\sqrt{N_0}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \approx \mathbb{N}_q(\mathbf{0}, \sigma^2 (\frac{1}{N_0} \tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P^\intercal} \tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P})^{-1})$$

$$\Rightarrow \sqrt{N_0}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \approx \mathbb{N}_q(\mathbf{0}, \sigma^2 N_0 (\tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P^\intercal} \tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P})^{-1})$$

$$\Rightarrow (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \approx \mathbb{N}_q(\mathbf{0}, \sigma^2 (\tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P^\intercal} \tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P})^{-1}), \tag{2.13}$$

with $\tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P} := \frac{\partial g_0^P(\boldsymbol{a}, \boldsymbol{x}_0^P)}{\partial \boldsymbol{a}}|_{\boldsymbol{a}=\boldsymbol{\alpha}}$. Since $\boldsymbol{\alpha}$ is unknown $\hat{\boldsymbol{\alpha}}$ is replaced in $\tilde{X}_{0,\boldsymbol{\alpha}}^{P}$.

2.2.2 Prediction method of Seber

In regression models, we are interested in predicting the future observation, Y_0^F, which is modeled as

$$Y_0^F = g(\boldsymbol{\alpha}, x_0^F) + E_0^F, \tag{2.14}$$

where $g(\boldsymbol{\alpha}, x_0^F)$ is a nonlinear function of $\boldsymbol{\alpha}$ and $E_0^F \sim \mathbb{N}(0, \sigma^2)$ such that E_0^P and E_0^F are independent. The prediction of the future observation is given by

$$\hat{Y}_0^F = g(\hat{\boldsymbol{\alpha}}, x_0^F),$$

where $\hat{\boldsymbol{\alpha}}$ is obtained from the Gauss-Newton method. This prediction can be obtained using the *predict* function in the stats package in R (R Development Core Team, 2016).

Theorem 2.4. *Prediction interval for Y_0^F [Seber (1989)]*
Consider the nonlinear regression model (2.8) for the partially observed individual and the nonlinear model (2.14) for the future observation from the new individual. Prediction of the future observation, Y_0^F, is given by

$$\hat{Y}_0^F = g(\hat{\boldsymbol{\alpha}}, x_0^F),$$

where $\hat{\boldsymbol{\alpha}}$ is obtained from the Gauss-Newton method. The prediction error $(\hat{Y}_0^F - Y_0^F)$ has normal distribution as follows:

$$\hat{Y}_0^F - Y_0^F \approx \mathbb{N}(0, \sigma^2[\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F(\tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P\intercal}\tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^P)^{-1}\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^{F\intercal} + 1]), \tag{2.15}$$

where $\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F = \frac{\partial g(\boldsymbol{a}, x_0^F)}{\partial \boldsymbol{a}}|_{\boldsymbol{a}=\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^P = \frac{\partial g_0^P(\boldsymbol{a}, \boldsymbol{x}_0^P)}{\partial \boldsymbol{a}}|_{\boldsymbol{a}=\boldsymbol{\alpha}}$.
Then the $(1-\alpha)$-prediction interval for Y_0^F is given by

$$PI = (\hat{y}_0^F \pm q_{\mathbb{N}(0,1),1-\alpha/2}\sqrt{\hat{\sigma}^2[\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^F(\tilde{\boldsymbol{X}}_{0,\hat{\boldsymbol{\alpha}}}^{P\intercal}\tilde{\boldsymbol{X}}_{0,\hat{\boldsymbol{\alpha}}}^P)^{-1}\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal} + 1]}), \tag{2.16}$$

where $\hat{y}_0^F$ is the realization of $\hat{Y}_0^F$.

Proof. The proof can be found in the appendix. □

3 Prediction using only observations of the other series

3.1 Prediction of the future observation in linear random effects models

Assume that the N_i observations of the ith individual are interpreted by the model

$$Y_i = \boldsymbol{X}_i A_i + E_i, \quad i = 1, \cdots, I, \tag{3.1}$$

where $Y_i = (Y_{i1}, \cdots, Y_{iN_i})^\intercal$ is the observation vector of the ith individual, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}^\intercal, \cdots, \boldsymbol{x}_{iN_i}^\intercal)^\intercal$ is the $N_i \times q$ covariate matrix of the ith individual, where $\boldsymbol{x}_{ij} := (1, h_1(x_{ij}), \cdots, h_{q-1}(x_{ij}))$, $\quad i \in \{1, \cdots, I\}$, $\quad j \in \{1, \cdots, N_i\}$, and A_i and E_i are respectively the $(q \times 1)$-vector of coefficients and the $(N_i \times 1)$-vector of errors. If A_i's are supposed to be independent and random then the model (3.1) is called a random coefficient regression model (RCR). In the case of balanced data ($N_i = N, \boldsymbol{X}_i = \boldsymbol{X}$, $\quad i = 1, \cdots, I$), Rao (1965) and Swamy (1971) present estimation procedures for a RCR model. Moreover, Carter and Yang (1986) and Johansen (1982) propose estimation methods for the unbalanced data. In the following, all assumptions of a RCR model for the balanced longitudinal data are presented.

Assumption 3.1. In the defined RCR model of (3.1) with the balanced data, the following assumptions hold (Swamy, 1971):
[A1] $\boldsymbol{X}$ is a nonstochastic $(N \times q)$-design matrix which includes the independent variables. The sample size N is larger than q $(N > q)$. In addition, $\boldsymbol{X}$ is full rank, i.e. $\text{rank}(\boldsymbol{X}) = q$.

[A2] Stochastic coefficient vectors A_i, $\quad i \in \{1, \cdots, I\}$ are independent and identically distributed with mean $\boldsymbol{\alpha} \in \mathbb{R}^q$ and covariance matrix $\Sigma \in \mathbb{R}^{q \times q}$. The stochastic character of the coefficient vectors can also be expressed as $A_i = \boldsymbol{\alpha} + \boldsymbol{\delta}_i$, where $\boldsymbol{\delta}_i := (\delta_{1i}, \cdots, \delta_{qi})^\intercal$ is a random vector with mean $\mathbf{0}$ and variance covariance Σ. The components of $\boldsymbol{\alpha} := (\alpha_1, \cdots, \alpha_q)^\intercal$ and $\boldsymbol{\delta}_i$ are respectively considered as the fixed (nonstochastic) and random effects.

[A3] Stochastic $(N \times 1)$-error vectors E_i, $\quad i \in \{1, \cdots, I\}$ are independent and identically distributed with mean $\mathbf{0}$ and variance covariance matrix $\sigma^2 \boldsymbol{I}_N$, where $\sigma^2 > 0$ and $\boldsymbol{I}_N$ is a $(N \times N)$-identity matrix.

[**A4**] Random vectors of A_i and E_j are stochastically independent for all $i, j \in \{1, \cdots, I\}$, i.e. $A_1, \cdots, A_I, E_1, \cdots, E_I$ are stochastically independent.

[**A5**] Both the coefficient and error vectors A_i and $E_i, \quad i \in \{1, \cdots, I\}$ are distributed as multivariate normal. Hence, for all $i \in \{1, \cdots, I\}$ we have $A_i \sim \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$ and $E_i \sim \mathbb{N}_N(\mathbf{0}, \sigma^2 \boldsymbol{I}_N)$.

The assumption of normal distribution [A5] is unnecessary to derive the parameter estimates, however it is essential to construct a prediction interval. According to the assumptions [A1]-[A5], the observation vectors $Y_1, \cdots, Y_I$ are independent and identically distributed as follows:

$$Y_1, \cdots, Y_I \sim \mathbb{N}_N(\boldsymbol{X}\boldsymbol{\alpha}, \boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \sigma^2 \boldsymbol{I}_N). \tag{3.2}$$

3.1.1 Estimation method of Swamy in the balanced case

In a RCR model, q fixed effects from the mean vector $\boldsymbol{\alpha}$, $\frac{q(q+1)}{2}$ components from the covariance matrix Σ and the intraindividual variance σ^2 are the unknown parameters to be estimated. In this section, the proposed estimation method of Swamy (1971) is described.

3.1.1.1 Estimation of fixed effects

For derivation of the best linear unbiased estimator of the parameter vector $\boldsymbol{\alpha}$, firstly the I regression equations $Y_i = \boldsymbol{X}A_i + E_i, \quad i \in \{1, \cdots, I\}$ are considered. Let $\hat{A}_{i,OLS} := (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal Y_i$ be the OLS-Estimation of the parameter vector A_i. Based on the assumptions [A1]-[A5], the distribution of OLS-Estimations $\hat{A}_{i,OLS}, \quad i \in \{1, \cdots, I\}$ is given by

$$\hat{A}_{1,OLS}, \cdots, \hat{A}_{I,OLS} \sim \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma + \sigma^2(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}). \tag{3.3}$$

A simple estimation of $\boldsymbol{\alpha}$ is the mean of OLS-Estimations $\hat{A}_{i,OLS}, \quad i \in \{1, \cdots, I\}$. In the following theorem, it will be proved that $\hat{\boldsymbol{\alpha}}_{GLS}$, which is defined as $\hat{\boldsymbol{\alpha}}_{GLS} := \frac{1}{I}\sum_{i=1}^{I} \hat{A}_{i,OLS}$ is the best linear unbiased estimation of $\boldsymbol{\alpha}$. For the proof of this theorem, we firstly need to find the inverse of $(\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \sigma^2 \boldsymbol{I}_N)$.

Lemma 3.1. *Inverse of* $(\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \sigma^2\boldsymbol{I}_N)$ *[Swamy (1971)]*
Let $\sigma^2 > 0$, $\boldsymbol{I}_N$ *be the* $(N \times N)$*-identity matrix,* $\Sigma \in \mathbb{R}^{q\times q}$ *be a nonsingular matrix and* $\boldsymbol{X}$ *the* $(N \times q)$ *matrix* $(N > q)$ *with rank*$(\boldsymbol{X}) = q$*. Then the inverse of* $(\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \sigma^2\boldsymbol{I}_N)$ *is given by*

$$\sigma^{-2}\boldsymbol{I}_N - \sigma^{-2}\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal + \boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}(\Sigma + \sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal. \tag{3.4}$$

Proof. The proof can be found in the appendix. □

Theorem 3.1. *Best linear unbiased estimation of fixed effects [Swamy (1971)]*
Consider the RCR model $Y_i = \boldsymbol{X}A_i + E_i, \quad i \in \{1, \cdots, I\}$ *and the assumptions [A1]-[A5]. The best linear unbiased estimation of* $\boldsymbol{\alpha}$ *is given by*

$$\hat{\boldsymbol{\alpha}}_{GLS} := \frac{1}{I}\sum_{i=1}^{I}\hat{A}_{i,OLS}.$$

Proof. The proof can be found in the appendix. □

From (3.3) and the independency between $\hat{A}_{1,OLS}, \cdots, \hat{A}_{I,OLS}$, we conclude that

$$\hat{\boldsymbol{\alpha}}_{GLS} \sim \mathbb{N}_q(\boldsymbol{\alpha}, \frac{1}{I}(\Sigma + \sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})). \tag{3.5}$$

3.1.1.2 Estimation of variance components

Based on the assumptions [A1]-[A5], Swamy (1971) proposes the unbiased and consistent estimations for the intraindividual variance, σ^2, and the covariance matrix, Σ.

Theorem 3.2. *Estimation of* σ^2 *and* Σ *[Swamy (1971)]*
Consider the RCR model $Y_i = \boldsymbol{X}A_i + E_i, \quad i \in \{1, \cdots, I\}$ *and the assumptions [A1]-[A5]. The unbiased and consistent estimations of* σ^2 *and* Σ *are given by*

$$\hat{\sigma}^2 := \frac{1}{I}\sum_{i=1}^{I}\hat{\sigma}_i^2 = \frac{1}{I}\sum_{i=1}^{I}\frac{(Y_i - \boldsymbol{X}\hat{A}_{i,OLS})^\intercal(Y_i - \boldsymbol{X}\hat{A}_{i,OLS})}{N-q}, \tag{3.6}$$

$$\hat{\Sigma} := \frac{1}{I-1}\sum_{i=1}^{I}(\hat{A}_{i,OLS} - \hat{\boldsymbol{\alpha}}_{GLS})(\hat{A}_{i,OLS} - \hat{\boldsymbol{\alpha}}_{GLS})^\intercal - \hat{\sigma}^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}, \tag{3.7}$$

where $\hat{\sigma}_i^2$ *is the variance estimation of the ith regression model* $Y_i = \boldsymbol{X}A_i + E_i$*.*

Proof. The proof can be found in the appendix. □

3.1.2 Prediction method of Swamy

In many applications, the prediction of the future observation is interested. We define $Y_0^F = \boldsymbol{x}_0^F A_0 + E_0^F$ as the unknown observation of a new individual at x_0^F, where $\boldsymbol{x}_0^F = (1, h_1(x_0^F), \cdots, h_{q-1}(x_0^F)) \in \mathbb{R}^q$. The new individual is also modeled as a RCR model. Hence, it is concluded that $A_0 \sim \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$, $E_0^F \sim \mathbb{N}(0, \sigma^2)$ and $E_0^F, E_1, \cdots, E_I, A_0, A_1, \cdots, A_I$ are stochastically independent. The best linear unbiased predictor for Y_0^F is given by (Swamy, 1971):

$$\hat{y}_0^F := \boldsymbol{x}_0^F \hat{\boldsymbol{\alpha}}_{GLS},$$

where $\hat{\boldsymbol{\alpha}}_{GLS} := \frac{1}{I}\sum_{i=1}^{I} \hat{a}_{i,OLS}$ and $\hat{a}_{i,OLS}$ is the realization of $\hat{A}_{i,OLS}$. This predictor is the best linear unbiased predictor for Y_0^F because, $\hat{\boldsymbol{\alpha}}_{GLS}$ is the best linear unbiased estimation of the fixed effects. In the Lemma 3.2 and Theorem 3.3, the distribution of prediction error $(\hat{Y}_0^F - Y_0^F)$ and a prediction interval for Y_0^F are derived by Swamy (1971).

Lemma 3.2. *Distribution of the prediction error [Swamy (1971)]*
Consider the RCR model $Y_i = \boldsymbol{X} A_i + E_i, \quad i \in \{1, \cdots, I\}$ and the assumptions [A1]-[A5]. Let $\hat{Y}_0^F := \boldsymbol{x}_0^F \hat{\boldsymbol{\alpha}}_{GLS}$ be the prediction for the observation of the new individual at x_0^F, where $\hat{\boldsymbol{\alpha}}_{GLS} := \frac{1}{I}\sum_{i=1}^{I} \hat{A}_{i,OLS}$. Then the prediction error $(\hat{Y}_0^F - Y_0^F)$ has normal distribution with the mean 0 and variance

$$\nu := \frac{1}{I}\boldsymbol{x}_0^F(\Sigma + \sigma^2(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal + \boldsymbol{x}_0^F \Sigma (\boldsymbol{x}_0^F)^\intercal + \sigma^2. \tag{3.8}$$

Proof. The proof can be found in the appendix. □

If the variance components Σ and σ^2 are known, then

$$\frac{\hat{Y}_0^F - Y_0^F}{\sqrt{\nu}} \sim \mathbb{N}(0, 1).$$

However, if the variance components are unknown the estimations of Σ and σ^2 are needed to construct an approximate prediction interval.

Theorem 3.3. *Approximate prediction interval for Y_0^F [Swamy (1971)]*
Consider the RCR model $Y_i = \boldsymbol{X} A_i + E_i, \quad i \in \{1, \cdots, I\}$ and the assumptions [A1]-[A5]. Let $\hat{Y}_0^F := \boldsymbol{x}_0^F \hat{\boldsymbol{\alpha}}_{GLS}$ be the prediction for the observation of the new individual at x_0^F, where $\hat{\boldsymbol{\alpha}}_{GLS} := \frac{1}{I}\sum_{i=1}^{I} \hat{A}_{i,OLS}$. Then the variance of prediction error ν can be estimated as

$$\hat{\nu} := \frac{1}{I}\boldsymbol{x}_0^F(\hat{\Sigma} + \hat{\sigma}^2(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal + \boldsymbol{x}_0^F \hat{\Sigma} (\boldsymbol{x}_0^F)^\intercal + \hat{\sigma}^2. \tag{3.9}$$

By use of $\hat{\nu}$, the standardized prediction error converges to the standard normal distribution for $N, I \to \infty$. Consequently, the $(1-\alpha)$-prediction interval for Y_0^F is given by

$$\hat{\mathcal{PI}}(\boldsymbol{y}) = [\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}], \tag{3.10}$$

where $\hat{y}_0^F$ is the realization of $\hat{Y}_0^F$ and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of standard normal distribution.

Proof. The proof can be found in the appendix. □

3.2 Prediction of the future observation in nonlinear random effects models

Nonlinear mixed effects (NLME) model includes the nonlinear function of some or all of the fixed and random effects. Lindstrom and Bates (1990) propose the NLME model for repeated measures as follows:

$$Y_{ij} = g(A_i, x_{ij}) + E_{ij}, \quad i = 1, \cdots, I, \quad j = 1, \cdots, N_i, \tag{3.11}$$

where Y_{ij} is the jth response on the ith individual (group), $g : \mathbb{R}^q \times \mathbb{R} \to \mathbb{R}$ is a nonlinear and differentiable function of the parameter vector A_i, x_{ij} is the independent variable, and E_{ij} is the within group error term which has normal distribution. The parameter vector A_i can be different from individual to individual and is modeled as

$$A_i = \boldsymbol{\alpha} + \boldsymbol{\delta}_i, \quad \boldsymbol{\delta}_i \overset{iid}{\sim} \mathbb{N}_q(\mathbf{0}, \Sigma), \tag{3.12}$$

where $\boldsymbol{\alpha}$ is a q-dimensional vector of fixed effects and $\boldsymbol{\delta}_i$ is a q-dimensional random effects vector related to the ith individual which has normal distribution with mean zero and the positive definite variance covariance matrix Σ. It is also supposed that the within group errors E_{ij} are distributed independently as $\mathbb{N}(0, \sigma^2)$ and independent of the $\boldsymbol{\delta}_i$ and that observations related to the different individuals are independent. For all $i \in \{1, \cdots, I\}$, the nonlinear random effects model is modeled as

$$Y_i = g_i(A_i, \boldsymbol{x}_i) + E_i, \quad g_i(A_i, \boldsymbol{x}_i) : \mathbb{R}^q \times \mathbb{R}^{N_i} \to \mathbb{R}^{N_i} \tag{3.13}$$

where A_i is a q-dimensional vector of random effects which is independently distributed as $\mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$ and

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{iN_i} \end{pmatrix}, \boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iN_i} \end{pmatrix}, g_i(A_i, \boldsymbol{x}_i) := \begin{pmatrix} g(A_i, x_{i1}) \\ \vdots \\ g(A_i, x_{iN_i}) \end{pmatrix}, E_i = \begin{pmatrix} E_{i1} \\ \vdots \\ E_{iN_i} \end{pmatrix}.$$

In this thesis we consider $g(A_i, x_{ij}) = A_{0i} + A_{1i} x_{ij}^{A_{2i}}, \quad j = 1, \cdots, N_i$.

3.2.1 Estimation method of Pinheiro and Bates in the unbalanced case

Various methods have been suggested for the estimation of parameters in the NLME models. Some of them involve taking a first-order Taylor expansion of g around the conditional mode of the random effects (Lindstrom and Bates, 1990), or around the expected value of random effects (Sheiner and Beal, 1980; Vonesh and Carter, 1992). In this section, the estimation procedure proposed by Lindstrom and Bates (1990) for the defined NLME model (3.13) will be described. In mixed models, because the random effects are not observed the marginal density of the response $\boldsymbol{Y}$ is applied for the maximum likelihood estimation of parameters which is calculated as

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \int f(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{\alpha}, \sigma^2) f(\boldsymbol{\delta}|\Sigma) d\boldsymbol{\delta}, \tag{3.14}$$

where $\boldsymbol{\beta} := (\boldsymbol{\alpha}, \sigma^2, \Sigma)$, $\boldsymbol{\delta} := (\boldsymbol{\delta}_1^\intercal, \cdots, \boldsymbol{\delta}_I^\intercal)^\intercal$ and $\boldsymbol{y}$ is the realization of $\boldsymbol{Y}$. For the NLME model (3.13), Pinheiro and Bates (2000) propose to express the variance covariance matrix of random effects by the unknown precision factor $\Delta \in \mathbb{R}^{q \times q}$, such that $\Sigma^{-1} = \sigma^{-2} \Delta^\intercal \Delta$ and Δ is represented by an unconstrained set of parameters $\boldsymbol{\theta}$. For a positive-definite matrix Σ, such a Δ is available however it does not need to be unique. By use of the unknown precision factor $\Delta(\boldsymbol{\theta})$, the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{\delta}$ can be written as

$$f(\boldsymbol{y}, \boldsymbol{\delta}|\boldsymbol{\alpha}, \sigma^2, \Delta(\boldsymbol{\theta})) = f(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{\alpha}, \sigma^2) f(\boldsymbol{\delta}|\Delta(\boldsymbol{\theta})) = \prod_{i=1}^{I} f(y_i|\boldsymbol{\delta}_i, \boldsymbol{\alpha}, \sigma^2) f(\boldsymbol{\delta}_i|\Delta(\boldsymbol{\theta}))$$

$$= \prod_{i=1}^{I} \frac{1}{\sqrt{(2\pi\sigma^2)^{N_i}}} \exp\{\frac{||y_i - g_i(\boldsymbol{\delta}_i + \boldsymbol{\alpha}, \boldsymbol{x}_i)||^2}{-2\sigma^2}\} \cdot \frac{\exp\{\frac{||\Delta(\boldsymbol{\theta})\boldsymbol{\delta}_i||^2}{-2\sigma^2}\}}{\sqrt{(2\pi)^q \cdot |\sigma^2 (\Delta(\boldsymbol{\theta})^\intercal \Delta(\boldsymbol{\theta}))^{-1}|}}$$

$$= \frac{|\Delta(\boldsymbol{\theta})|^I}{\sqrt{(2\pi\sigma^2)^{(N+Iq)}}}\exp\{\frac{\sum_{i=1}^I[||y_i - g_i(\boldsymbol{\delta}_i+\boldsymbol{\alpha},\boldsymbol{x}_i)||^2 + ||\Delta(\boldsymbol{\theta})\boldsymbol{\delta}_i||^2]}{-2\sigma^2}\}, \tag{3.15}$$

where $N := \sum_{i=1}^I N_i$. The alternating estimation algorithm suggested by Lindstrom and Bates (1990) has two steps, a penalized nonlinear least squares (PNLS) step and a linear mixed effects (LME) step.

In the PNLS step, at the $(k+1)$th iteration, Lindstrom and Bates (1990) mention that the maximum likelihood estimations of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}_i$ maximize (3.15) by given $\boldsymbol{\theta}^{(k)}$ or equivalently minimize the following penalized nonlinear least squares function

$$\sum_{i=1}^I[||y_i - g_i(\boldsymbol{\delta}_i+\boldsymbol{\alpha},\boldsymbol{x}_i)||^2 + ||\Delta(\boldsymbol{\theta}^{(k)})\boldsymbol{\delta}_i||^2]. \tag{3.16}$$

In the LME step, at the $(k+1)$th iteration the estimate of $\Delta(\boldsymbol{\theta})$ and σ^2 are updated based on the first order Taylor expansion of $g_i(\boldsymbol{\delta}_i+\boldsymbol{\alpha},\boldsymbol{x}_i)$ around $\hat{\boldsymbol{\delta}}_i^{(k+1)}$ and $\hat{\boldsymbol{\alpha}}^{(k+1)}$, which are obtained from the PNLS step. Let

$$\tilde{\boldsymbol{X}}_i^{(k+1)} := \frac{\partial g_i(\boldsymbol{a},\boldsymbol{x}_i)}{\partial \boldsymbol{a}^\intercal}|_{\boldsymbol{a}=\hat{\boldsymbol{\alpha}}^{(k+1)}+\hat{\boldsymbol{\delta}}_i^{(k+1)}} \in \mathbb{R}^{N_i\times q},$$

$$\hat{Z}_i^{(k+1)} := Y_i - g_i(\hat{\boldsymbol{\delta}}_i^{(k+1)}+\hat{\boldsymbol{\alpha}}^{(k+1)},\boldsymbol{x}_i)+\tilde{\boldsymbol{X}}_i^{(k+1)}(\hat{\boldsymbol{\delta}}_i^{(k+1)}+\hat{\boldsymbol{\alpha}}^{(k+1)}) \approx \tilde{\boldsymbol{X}}_i^{(k+1)}(\boldsymbol{\delta}_i+\boldsymbol{\alpha})+E_i. \tag{3.17}$$

The applied approximate log-likelihood function for estimation of $\Delta(\boldsymbol{\theta})$ and σ^2 at the $(k+1)$th iteration is given by

$$l_{LME}(\hat{\boldsymbol{\alpha}}^{(k+1)},\sigma^2,\Delta(\boldsymbol{\theta})|\boldsymbol{y}) \approx -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^I\{\log|V_i^{(k+1)}(\boldsymbol{\theta})|$$

$$+\sigma^{-2}[\hat{z}_i^{(k+1)} - \tilde{\boldsymbol{X}}_i^{(k+1)}\hat{\boldsymbol{\alpha}}^{(k+1)}]^\intercal V_i^{(k+1)^{-1}}(\boldsymbol{\theta})[\hat{z}_i^{(k+1)} - \tilde{\boldsymbol{X}}_i^{(k+1)}\hat{\boldsymbol{\alpha}}^{(k+1)}]\}, \tag{3.18}$$

where $V_i^{(k+1)}(\boldsymbol{\theta}) := \boldsymbol{I}_{N_i}+\tilde{\boldsymbol{X}}_i^{(k+1)}\Delta(\boldsymbol{\theta})^{-1}\Delta(\boldsymbol{\theta})^{-\intercal}\tilde{\boldsymbol{X}}_i^{(k+1)^\intercal}$ and $\hat{z}_i^{(k+1)}$ is the realization of $\hat{Z}_i^{(k+1)}$. This log-likelihood is related to the log-likelihood of a linear random effects model such that $\hat{\boldsymbol{Z}}^{(k+1)} := (\hat{Z}_1^{(k+1)^\intercal},\cdots,\hat{Z}_I^{(k+1)^\intercal})^\intercal$is the response vector and $\tilde{\boldsymbol{X}}^{(k+1)} := (\tilde{\boldsymbol{X}}_1^{(k+1)^\intercal},\cdots,\tilde{\boldsymbol{X}}_I^{(k+1)^\intercal})^\intercal$ is the fixed- and random-effects design matrix.

The algorithm iterates between the PNLS and LME steps till convergence. The

computational methods of this algorithm have been presented in the appendix. By use of the *nlme* function of package nlme of Pinheiro et al. (2016), the estimates using this procedure can be obtained.
Lindstrom and Bates (1990) only apply the LME step for updating the estimates of $\Delta(\boldsymbol{\theta})$ and σ^2. Whereas, Wolfinger (1993) iterates LME steps by recalculating (3.17) and (3.18) at the updated estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}_i$.

3.2.1.1 Inference Results

Under the LME approximation and certain regularity conditions which are generally satisfied in practice, Pinheiro (1994) proves that the maximum likelihood estimations in the LME model are consistent and asymptotically normal. Therefore, $\hat{\boldsymbol{\alpha}}$ has approximately normal distribution with mean $\boldsymbol{\alpha}$ and the estimated variance covariance matrix $\widehat{\mathbb{C}\text{ov}}(\hat{\boldsymbol{\alpha}})$ as follows:

$$\mathbb{E}(\hat{\boldsymbol{\alpha}}) \approx \boldsymbol{\alpha}, \quad \widehat{\mathbb{C}\text{ov}}(\hat{\boldsymbol{\alpha}}) \approx [\sum_{i=1}^{I} \tilde{\boldsymbol{X}}_i^{\intercal} \hat{V}_i^{-1} \tilde{\boldsymbol{X}}_i]^{-1}, \tag{3.19}$$

where

$$\hat{V}_i := \tilde{\boldsymbol{X}}_i \Sigma(\hat{\boldsymbol{\theta}}) \tilde{\boldsymbol{X}}_i^{\intercal} + \hat{\sigma}^2 \boldsymbol{I}_{N_i}, \quad \Sigma(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2 (\Delta^{\intercal}(\hat{\boldsymbol{\theta}}) \Delta(\hat{\boldsymbol{\theta}}))^{-1}$$

and

$$\tilde{\boldsymbol{X}}_i := \frac{\partial g_i(\boldsymbol{a}, \boldsymbol{x}_i)}{\partial \boldsymbol{a}^{\intercal}}|_{\boldsymbol{a}=\hat{\boldsymbol{\alpha}}+\hat{\boldsymbol{\delta}}_i}.$$

Moreover, $(\hat{\boldsymbol{\theta}}, \log(\hat{\sigma}))$ has approximately normal distribution as follows:

$$\begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \log \hat{\sigma} \end{pmatrix} \sim \mathbb{N}(\begin{pmatrix} \boldsymbol{\theta} \\ \log \sigma \end{pmatrix}, \mathbb{I}^{-1}(\boldsymbol{\theta}, \sigma)), \tag{3.20}$$

where

$$\mathbb{I} = - \begin{pmatrix} \partial^2 l_{LME_p} / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\intercal} & \partial^2 l_{LME_p} / \partial \log \sigma \partial \boldsymbol{\theta}^{\intercal} \\ \partial^2 l_{LME_p} / \partial \boldsymbol{\theta} \partial \log \sigma & \partial^2 l_{LME_p} / \partial^2 \log \sigma \end{pmatrix},$$

and l_{LME_p} denotes the profiled LME approximation to the log-likelihood on the fixed effects. Practically, σ^2 and $\boldsymbol{\theta}$ are unknown hence their estimates should be replaced in (3.20). Since the inference results for NLME models are based on the LME approximation to the log-likelihood, they are less reliable than asymptotic inferences for LME models (Pinheiro and Bates, 2000).

3.2.2 Extension of the prediction method of Pinheiro and Bates (ePB)

Pinheiro and Bates (2000) propose a simple plug-in predictor by estimating the expected response when the random effects are equal to their mean value $\mathbf{0}$. We define $Y_0^F = g(A_0, x_0^F) + E_0^F$ as the unknown observation of a new individual at x_0^F. The new individual is also modeled as a NLME model. Hence, it is concluded that $A_0 := \boldsymbol{\alpha} + \boldsymbol{\delta}_0 \sim \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$, $E_0^F \sim \mathbb{N}(0, \sigma^2)$ and $E_0^F, E_1, \cdots, E_I, A_0, A_1, \cdots, A_I$ are stochastically independent. The proposed simple plug-in prediction of Y_0^F when $\boldsymbol{\delta}_0 = \mathbf{0}$ is given by

$$\hat{Y}_0^F := g(\hat{\boldsymbol{\alpha}}, x_0^F), \tag{3.21}$$

where $\hat{\boldsymbol{\alpha}}$, the maximum likelihood estimation of $\boldsymbol{\alpha}$, is obtained using the proposed estimation procedure by Pinheiro and Bates (2000) which has been described in details in the appendix (the *predict* function of the package nlme of Pinheiro et al. (2016) gives this point prediction). In the following theorem, by use of the method for getting prediction interval in nonlinear regression models we suggest an approximation to get the approximate $(1-\alpha)$-prediction interval for Y_0^F based on the Pinheiro and Bates' prediction.

Theorem 3.4. *Approximate prediction interval for Y_0^F based on the Pinheiro and Bates' prediction:*
Consider the following nonlinear random effects model

$$Y_i = g_i(A_i, \boldsymbol{x}_i) + E_i, \quad i \in \{1, \cdots, I\}$$

where

$$A_i = \boldsymbol{\alpha} + \boldsymbol{\delta}_i \quad \boldsymbol{\delta}_i \overset{iid}{\sim} \mathbb{N}_q(\mathbf{0}, \Sigma), \quad E_i \overset{iid}{\sim} \mathbb{N}_{N_i}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{N_i}).$$

Let the new observation of the new individual be modeled as $Y_0^F = g(A_0, x_0^F) + E_0^F$, where $A_0 = \boldsymbol{\alpha} + \boldsymbol{\delta}_0$, $\boldsymbol{\delta}_0 \sim \mathbb{N}_q(\mathbf{0}, \Sigma)$ and $E_0^F \sim \mathbb{N}(0, \sigma^2)$. Assume $A_0, A_1, \cdots, A_I$, $E_0^F, E_1, \cdots, E_I$ are independent. Let a simple plug-in predictor of Y_0^F at x_0^F be $\hat{Y}_0^F := g(\hat{\boldsymbol{\alpha}}, x_0^F)$. Then the prediction error has approximately normal distribution with mean 0 and the estimated variance

$$\hat{\nu} := \tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^F [\widehat{\mathbb{C}ov}(\hat{\boldsymbol{\alpha}}) + \hat{\Sigma}] \tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal} + \hat{\sigma}^2,$$

where $\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^F := \frac{\partial g(\boldsymbol{a}, x_0^F)}{\partial \boldsymbol{a}^\intercal}|_{\boldsymbol{a}=\hat{\boldsymbol{\alpha}}}$ and $\widehat{\mathbb{C}ov}(\hat{\boldsymbol{\alpha}}) \approx [\sum_{i=1}^I \tilde{\boldsymbol{X}}_i^\intercal \hat{V}_i^{-1} \tilde{\boldsymbol{X}}_i]^{-1}$ with $\hat{V}_i := \tilde{\boldsymbol{X}}_i \hat{\Sigma} \tilde{\boldsymbol{X}}_i^\intercal + \hat{\sigma}^2 \boldsymbol{I}_{N_i}$, $\tilde{\boldsymbol{X}}_i := \frac{\partial g_i(\boldsymbol{a}, \boldsymbol{x}_i)}{\partial \boldsymbol{a}^\intercal}|_{\boldsymbol{a}=\hat{\boldsymbol{\alpha}}+\hat{\boldsymbol{\delta}}_i}$ and $\hat{\Sigma} = \hat{\sigma}^2 (\Delta^\intercal(\hat{\boldsymbol{\theta}}) \Delta(\hat{\boldsymbol{\theta}}))^{-1}$ with $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\delta}}_i$ which are the estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$ and $\boldsymbol{\delta}_i$ at the final iteration of the Pinheiro and Bates' estimation method. Moreover, the approximate $(1-\alpha)$-prediction interval for Y_0^F from a new individual is given by

$$[\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}], \tag{3.22}$$

where $\hat{y}_0^F$ is the realization of $\hat{Y}_0^F$ and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of standard normal distribution.

Proof. If we consider the first order Taylor expansion of $g(A_0, x_0^F)$ and $g(\hat{\boldsymbol{\alpha}}, x_0^F)$ around $\boldsymbol{\alpha}$, then the prediction error can be simplified as

$$\begin{aligned} Y_0^F - \hat{Y}_0^F &= g(A_0, x_0^F) + E_0^F - g(\hat{\boldsymbol{\alpha}}, x_0^F) \\ &\approx g(\boldsymbol{\alpha}, x_0^F) + \tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F \boldsymbol{\delta}_0 - g(\boldsymbol{\alpha}, x_0^F) - \tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + E_0^F \\ &= \tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F \boldsymbol{\delta}_0 + E_0^F, \end{aligned} \tag{3.23}$$

where $\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F := \frac{\partial g(\boldsymbol{a}, x_0^F)}{\partial \boldsymbol{a}^\intercal}|_{\boldsymbol{a}=\boldsymbol{\alpha}}$. Since $\hat{\boldsymbol{\alpha}}$ is obtained based on the information of $\boldsymbol{Y} = (Y_1^\intercal, \cdots, Y_I^\intercal)^\intercal$ and $\boldsymbol{Y}$ is independent of E_0^F and $\boldsymbol{\delta}_0$ hence $\hat{\boldsymbol{\alpha}}$, E_0^F and $\boldsymbol{\delta}_0$ are also independent. Now, from (3.19) because $Y_0^F - \hat{Y}_0^F$ is approximately a linear combination of the three independent normal variables $(\hat{\boldsymbol{\alpha}}, \boldsymbol{\delta}_0, E_0^F)$, therefore we have

$$Y_0^F - \hat{Y}_0^F \sim \mathbb{N}(\boldsymbol{0}, \nu), \tag{3.24}$$

where

$$\nu = \tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F [\mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\alpha}}) + \Sigma] \tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^{F\intercal} + \sigma^2.$$

By replacing the ML estimates $\hat{\sigma}^2$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\alpha}}$, from the Pinheiro and Bates' estimation method, in ν the estimation $\hat{\nu}$ is given by

$$\hat{\nu} := \tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^F [\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}}) + \hat{\Sigma}] \tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal} + \hat{\sigma}^2.$$

Then the approximate $(1-\alpha)$-prediction interval for Y_0^F from a new individual is given by

$$[\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}],$$

where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of standard normal distribution. □

4 Prediction using all observations

4.1 Prediction of the future observation in linear random effects models

4.1.1 The modified method of Rao in the balanced case (mRao)

Let the defined RCR model (3.1) with the assumptions [A1]-[A5] for the balanced data. The goal is to predict the future observation from the new individual. Define $Y_0 = \boldsymbol{X}_0 A_0 + E_0$ as the observations of a new individual which is partitioned as

$$Y_0 = \begin{pmatrix} Y_0^P \\ Y_0^F \end{pmatrix}, \quad \boldsymbol{X}_0 = \begin{pmatrix} \boldsymbol{X}_0^P \\ \boldsymbol{x}_0^F \end{pmatrix} \quad \text{and} \quad E_0 = \begin{pmatrix} E_0^P \\ E_0^F \end{pmatrix},$$

where indices P and F show respectively the observed (Past) and unobserved (Future) components of the new individual and $\boldsymbol{X}_0^P = (\boldsymbol{x}_{01}^\intercal, \cdots, \boldsymbol{x}_{0N_0}^\intercal)^\intercal$ with $\boldsymbol{x_{0j}} = (1, h_1(x_{0j}), \cdots, h_{q-1}(x_{0j}))$ and $\boldsymbol{x}_0^F = (1, h_1(x_0^F), \cdots, h_{q-1}(x_0^F))$. The new individual is also modeled as a RCR model. Hence, it is concluded that $A_0 := \boldsymbol{\alpha} + \boldsymbol{\delta}_0 \sim \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$, $E_0 \sim \mathbb{N}_{N_0+1}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{N_0+1})$ and $E_0^P, E_0^F, E_1, \cdots, E_I, A_0, A_1, \cdots, A_I$ are stochastically independent.
Rao (1975) proposes the following prediction procedure for Y_0^F,

$$\hat{Y}_0^F = \boldsymbol{x}_0^F (\hat{C} \hat{A}_{0,OLS} + (\boldsymbol{I}_q - \hat{C}) \hat{\boldsymbol{\alpha}}_{GLS}), \tag{4.1}$$

where $\hat{C} := \hat{\Sigma}(\hat{\Sigma} + \hat{\sigma}^2 (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1})^{-1}$ and the estimates $\hat{\boldsymbol{\alpha}}_{GLS}$, $\hat{\Sigma}$ and $\hat{\sigma}^2$ are obtained based on the information of $\boldsymbol{Y} := (Y_1^\intercal, \cdots, Y_I^\intercal)^\intercal$ by use of the Swamy's estimation method and $\hat{A}_{0,OLS} := (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{X}_0^{P\intercal} Y_0^P$, which is computed only based on the information of Y_0^P. Like (3.3), it can be proved that

$$\hat{A}_{0,OLS} \sim \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma + \sigma^2 (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}). \tag{4.2}$$

A modification of Rao's prediction is given by supposing that the estimates of $\boldsymbol{\alpha}$, Σ and σ^2 are obtained based on the information of $\boldsymbol{Y}_{obs} := (Y_1^\intercal, \cdots, Y_I^\intercal, Y_0^{P\intercal})^\intercal$ as follows:

$$\hat{\boldsymbol{\alpha}}_{obs} = \frac{1}{I+1} \sum_{i=0}^{I} \hat{A}_{i,OLS}, \quad \hat{\boldsymbol{\alpha}}_{obs} \sim \mathbb{N}_q(\boldsymbol{\alpha}, v), \tag{4.3}$$

with

$$v = \frac{1}{(I+1)^2} [(I+1)\Sigma + \sigma^2 (I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} + (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1})],$$

and

$$\hat{\sigma}^2_{obs} := \frac{1}{I+1}[\sum_{i=1}^{I} \hat{\sigma}_i^2 + \hat{\sigma}_0^2]$$

$$= \frac{1}{I+1}[\sum_{i=1}^{I} \frac{(Y_i - \boldsymbol{X}\hat{A}_{i,OLS})^\intercal (Y_i - \boldsymbol{X}\hat{A}_{i,OLS})}{N-q}$$

$$+ \frac{(Y_0^P - \boldsymbol{X}_0^P \hat{A}_{0,OLS})^\intercal (Y_0^P - \boldsymbol{X}_0^P \hat{A}_{0,OLS})}{N_0 - q}],$$

$$\hat{\Sigma}_{obs} := \frac{1}{I}\sum_{i=0}^{I} (\hat{A}_{i,OLS} - \hat{\boldsymbol{\alpha}}_{obs})(\hat{A}_{i,OLS} - \hat{\boldsymbol{\alpha}}_{obs})^\intercal - \frac{\hat{\sigma}^2_{obs}}{I+1}[I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} + (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}],$$

where $\hat{A}_{i,OLS} := (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal Y_i$ and $\hat{A}_{0,OLS} := (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{X}_0^{P\intercal} Y_0^P$. Based on the unbiasedness and consistency of $\hat{\sigma}_i^2, \quad i = 0, \cdots, I$, the unbiasedness and consistency of $\hat{\sigma}^2_{obs}$ is proved. For studying the unbiasedness of $\hat{\Sigma}_{obs}$, we apply the used method in the proof of Theorem 3.2 for the expectation of $\hat{\Sigma}$ and we have

$$\mathbb{E}(\hat{\Sigma}_{obs}) = \frac{1}{I}[\sum_{i=1}^{I} \mathbb{E}(\hat{A}_{i,OLS} - \hat{\boldsymbol{\alpha}}_{obs})(\hat{A}_{i,OLS} - \hat{\boldsymbol{\alpha}}_{obs})^\intercal + \mathbb{E}(\hat{A}_{0,OLS} - \hat{\boldsymbol{\alpha}}_{obs})(\hat{A}_{0,OLS} - \hat{\boldsymbol{\alpha}}_{obs})^\intercal]$$

$$- \frac{\mathbb{E}(\hat{\sigma}^2_{obs})}{I+1}[I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} + (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}]$$

$$= \frac{1}{I}[I\mathbb{C}\mathrm{ov}(\hat{A}_{i,OLS}) - I\mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\alpha}}_{obs}) + \mathbb{C}\mathrm{ov}(\hat{A}_{0,OLS}) - \mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\alpha}}_{obs})]$$

$$- \frac{\sigma^2}{I+1}[I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} + (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}]$$

$$= \frac{1}{I}[I(\Sigma + \sigma^2(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}) - \frac{1}{I+1}((I+1)\Sigma + \sigma^2 I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} + \sigma^2(\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1})$$

$$+ \Sigma + \sigma^2(\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}] - \frac{\sigma^2}{I+1}[I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} + (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}]$$

$$= \frac{1}{I}[I\Sigma + \sigma^2 I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} - \Sigma - \frac{\sigma^2}{I+1} I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} - \frac{\sigma^2}{I+1}(\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}$$

$$+ \Sigma + \sigma^2(\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}] - \frac{\sigma^2}{I+1} I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} - \frac{\sigma^2}{I+1}(\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}$$

$$= \Sigma + \frac{\sigma^2}{I+1} I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} + \frac{\sigma^2}{I+1}(\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} - \frac{\sigma^2}{I+1} I(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} - \frac{\sigma^2}{I+1}(\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1}$$

$= \Sigma$.

Hence, $\hat{\Sigma}_{obs}$ is also an unbiased estimation of Σ. Like the proof of consistency of $\hat{\Sigma}$ in Theorem 3.2, by assuming that $B_1 := \lim_{N\to\infty}(N(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})$ and $B_2 := \lim_{N_0\to\infty}(N_0(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})$ exist and are positive definite it is readly concluded that $\hat{\Sigma}_{obs}$ is a consistent estimation of Σ and similar to (7.9) we have

$$\hat{\Sigma}_{obs} \xrightarrow{N,N_0\to\infty} \frac{S_{obs}}{I} \xrightarrow{I\to\infty} \Sigma, \tag{4.4}$$

where $S_{obs} := \sum_{i=0}^{I}(A_i - \bar{A})(A_i - \bar{A})^\intercal$, where $\bar{A}$ is the mean of A_i's.
In the following lemma and theorem, we calculate the distribution of prediction error $(\hat{Y}_0^F - Y_0^F)$ and propose an approximate prediction interval for Y_0^F based on the mentioned prediction procedure.

Lemma 4.1. *Distribution of the prediction error based on the modified Rao's prediction:*
Consider the RCR model $Y_i = \boldsymbol{X}A_i + E_i, \quad i \in \{1,\cdots,I\}$, and the assumptions [A1]-[A5]. Let the observations of the new individual be modeled as the RCR model $Y_0 = \boldsymbol{X}_0A_0 + E_0$ with the following decompositions

$$Y_0 = \begin{pmatrix} Y_0^P \\ Y_0^F \end{pmatrix}, \quad \boldsymbol{X}_0 = \begin{pmatrix} \boldsymbol{X}_0^P \\ \boldsymbol{x}_0^F \end{pmatrix} \quad \textit{and} \quad E_0 = \begin{pmatrix} E_0^P \\ E_0^F \end{pmatrix},$$

where indices P and F show respectively the observed (Past) and unobserved (Future) components of the new individual and $\boldsymbol{X}_0^P = (\boldsymbol{x}_{01}^\intercal,\cdots,\boldsymbol{x}_{0N_0}^\intercal)^\intercal$ with $\boldsymbol{x_{0j}} = (1, h_1(x_{0j}),\cdots,h_{q-1}(x_{0j}))$ and $\boldsymbol{x}_0^F = (1, h_1(x_0^F),\cdots,h_{q-1}(x_0^F))$. Hence, $A_0 \sim \mathbb{N}_q(\boldsymbol{\alpha},\Sigma)$, $E_0 \sim \mathbb{N}_{N_0+1}(\boldsymbol{0},\sigma^2\boldsymbol{I}_{N_0+1})$. Moreover, suppose $A_0, A_1,\cdots,A_I, E_0^P, E_0^F, E_1,\cdots,E_I$ are independent. If the variance components are known then the prediction for the new observation $Y_0^F = \boldsymbol{x}_0^FA_0 + E_0^F$ at x_0^F is given by

$$\hat{Y}_0^F := \boldsymbol{x}_0^F(C\hat{A}_{0,OLS} + (\boldsymbol{I}_q - C)\hat{\boldsymbol{\alpha}}_{obs}),$$

where

$$C := \Sigma[\Sigma + \sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})]^{-1}, \quad \hat{\boldsymbol{\alpha}}_{obs} = \frac{1}{I+1}\sum_{i=0}^{I}\hat{A}_{i,OLS}$$

and $\hat{A}_{i,OLS} := (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal Y_i$, $i \in \{1,\cdots,I\}$ and $\hat{A}_{0,OLS} := (\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}Y_0^P$. The prediction error has normal distribution with mean 0 and variance

$$\nu := \boldsymbol{x}_0^F[C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal + v - Cv - vC^\intercal + CvC^\intercal$$

$$+\frac{1}{I+1}C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}) - \frac{2}{I+1}C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal$$

$$+\frac{1}{I+1}(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal](\boldsymbol{x}_0^F)^\intercal+\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal+\sigma^2-2\boldsymbol{x}_0^F[C+\frac{(\boldsymbol{I}_q-C)}{I+1}]\Sigma(\boldsymbol{x}_0^F)^\intercal,$$

where v has been defined in (4.3).

Proof. Because $\hat{A}_{0,OLS}$ and $\hat{\boldsymbol{\alpha}}_{obs}$ are respectively linear in Y_0^P and $\boldsymbol{Y}_{obs}$, therefore $\hat{Y}_0^F$ is also linear in $\boldsymbol{Y}_{obs}$ and has normal distribution. Consequently, the prediction error has also normal distribution. For the mean of prediction error we have

$$\mathbb{E}(\hat{Y}_0^F-Y_0^F)=\mathbb{E}(\hat{Y}_0^F)-\mathbb{E}(Y_0^F)=\boldsymbol{x}_0^F[C\mathbb{E}(\hat{A}_{0,OLS})+(\boldsymbol{I}_q-C)\mathbb{E}(\hat{\boldsymbol{\alpha}}_{obs})]-\boldsymbol{x}_0^F\boldsymbol{\alpha}$$

$$=0. \quad \text{(from (4.2) and (4.3))}$$

For calculating the variance of prediction error, at first the variance of $\hat{Y}_0^F$ and the covariance of $\hat{Y}_0^F$ and Y_0^F are calculated as follows:

$$\mathbb{V}\text{ar}(\hat{Y}_0^F)=\boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS})C^\intercal+(\boldsymbol{I}_q-C)\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs})(\boldsymbol{I}_q-C)^\intercal$$

$$+C\mathbb{C}\text{ov}(\hat{A}_{0,OLS},\hat{\boldsymbol{\alpha}}_{obs})(\boldsymbol{I}_q-C)^\intercal+(\boldsymbol{I}_q-C)\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs},\hat{A}_{0,OLS})C^\intercal](\boldsymbol{x}_0^F)^\intercal$$

$$=\boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS})C^\intercal+\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs})-C\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs})-\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs})C^\intercal+C\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs})C^\intercal$$

$$+C\mathbb{C}\text{ov}(\hat{A}_{0,OLS},\hat{\boldsymbol{\alpha}}_{obs})-C\mathbb{C}\text{ov}(\hat{A}_{0,OLS},\hat{\boldsymbol{\alpha}}_{obs})C^\intercal+\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs},\hat{A}_{0,OLS})C^\intercal$$

$$-C\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs},\hat{A}_{0,OLS})C^\intercal](\boldsymbol{x}_0^F)^\intercal.$$

Since $\hat{A}_{0,OLS},\hat{A}_{1,OLS}\cdots,\hat{A}_{I,OLS}$ are linear functions of $Y_0^P,Y_1,\cdots,Y_I$ and $Y_0^P,Y_1,\cdots,Y_I$ are independent, hence $\hat{A}_{0,OLS},\hat{A}_{1,OLS},\cdots,\hat{A}_{I,OLS}$ are also independent. Now by use of this independency, $\mathbb{C}\text{ov}(\hat{A}_{0,OLS},\hat{\boldsymbol{\alpha}}_{obs})$ is given by

$$\mathbb{C}\text{ov}(\hat{A}_{0,OLS},\hat{\boldsymbol{\alpha}}_{obs})=\mathbb{C}\text{ov}(\hat{A}_{0,OLS},\frac{1}{I+1}\sum_{i=0}^{I}\hat{A}_{i,OLS})=\frac{1}{I+1}\mathbb{C}\text{ov}(\hat{A}_{0,OLS},\hat{A}_{0,OLS})$$

$$=\frac{1}{I+1}\mathbb{C}\text{ov}(\hat{A}_{0,OLS})=\frac{1}{I+1}(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}). \tag{4.5}$$

Therefore, by use of (4.3) and (4.5) we have

$$\mathbb{V}\text{ar}(\hat{Y}_0^F)=\boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS})C^\intercal+v-Cv-vC^\intercal+CvC^\intercal+\frac{1}{I+1}C\mathbb{C}\text{ov}(\hat{A}_{0,OLS})$$

$$-\frac{1}{I+1}C\mathbb{C}\text{ov}(\hat{A}_{0,OLS})C^\intercal+\frac{1}{I+1}\mathbb{C}\text{ov}(\hat{A}_{0,OLS})C^\intercal-\frac{1}{I+1}C\mathbb{C}\text{ov}(\hat{A}_{0,OLS})C^\intercal](\boldsymbol{x}_0^F)^\intercal$$

$$=\boldsymbol{x}_0^F[C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal+v-Cv-vC^\intercal+CvC^\intercal+\frac{1}{I+1}C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})$$

$$-\frac{2}{I+1}C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal+\frac{1}{I+1}(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal](\boldsymbol{x}_0^F)^\intercal$$

The $\mathbb{C}\text{ov}(\hat{Y}_0^F, Y_0^F)$ is calculated as

$$\begin{aligned}
\mathbb{C}\text{ov}(\hat{Y}_0^F, Y_0^F) &= \mathbb{C}\text{ov}(\boldsymbol{x}_0^F[C\hat{A}_{0,OLS}+(\boldsymbol{I}_q-C)\hat{\boldsymbol{\alpha}}_{obs}], \boldsymbol{x}_0^F A_0+E_0^F) \\
&= \mathbb{C}\text{ov}[\boldsymbol{x}_0^F(C\hat{A}_{0,OLS}+(\boldsymbol{I}_q-C)\hat{\boldsymbol{\alpha}}_{obs}), \boldsymbol{x}_0^F A_0] \\
&+\mathbb{C}\text{ov}[\boldsymbol{x}_0^F(C\hat{A}_{0,OLS}+(\boldsymbol{I}_q-C)\hat{\boldsymbol{\alpha}}_{obs}), E_0^F] \\
&= \boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS}, A_0)+(\boldsymbol{I}_q-C)\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs}, A_0)](\boldsymbol{x}_0^F)^\intercal \\
&+\boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS}, E_0^F)+(\boldsymbol{I}_q-C)\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs}, E_0^F)],
\end{aligned}$$

by replacing $\hat{A}_{0,OLS} = (\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}Y_0^P$ and based on the independency between $\hat{\boldsymbol{\alpha}}_{obs}$ and E_0^F and between $\hat{A}_{0,OLS}$ and E_0^F we have

$$\begin{aligned}
&\mathbb{C}\text{ov}(\hat{Y}_0^F, Y_0^F) = \boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS}, A_0)+(\boldsymbol{I}_q-C)\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{obs}, A_0)](\boldsymbol{x}_0^F)^\intercal \\
&= \boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS}, A_0)+\frac{(\boldsymbol{I}_q-C)}{I+1}\mathbb{C}\text{ov}(\hat{A}_{0,OLS}, A_0)](\boldsymbol{x}_0^F)^\intercal \\
&= \boldsymbol{x}_0^F[C+\frac{(\boldsymbol{I}_q-C)}{I+1}]\mathbb{C}\text{ov}(\hat{A}_{0,OLS}, A_0)(\boldsymbol{x}_0^F)^\intercal \\
&= \boldsymbol{x}_0^F[C+\frac{(\boldsymbol{I}_q-C)}{I+1}](\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}\mathbb{C}\text{ov}(Y_0^P, A_0)(\boldsymbol{x}_0^F)^\intercal \\
&= \boldsymbol{x}_0^F[C+\frac{(\boldsymbol{I}_q-C)}{I+1}](\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}\mathbb{C}\text{ov}(\boldsymbol{X}_0^P A_0+E_0^P, A_0)(\boldsymbol{x}_0^F)^\intercal \\
&= \boldsymbol{x}_0^F[C+\frac{(\boldsymbol{I}_q-C)}{I+1}](\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P\Sigma(\boldsymbol{x}_0^F)^\intercal \\
&= \boldsymbol{x}_0^F[C+\frac{(\boldsymbol{I}_q-C)}{I+1}]\Sigma(\boldsymbol{x}_0^F)^\intercal.
\end{aligned} \tag{4.6}$$

Finally, the variance of prediction error can be given by

$$\begin{aligned}
&\mathbb{V}\text{ar}(\hat{Y}_0^F-Y_0^F) = \mathbb{V}\text{ar}(\hat{Y}_0^F)+\mathbb{V}\text{ar}(Y_0^F)-2\mathbb{C}\text{ov}(\hat{Y}_0^F, Y_0^F) \\
&= \boldsymbol{x}_0^F[C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal+v-Cv-vC^\intercal+CvC^\intercal \\
&+\frac{1}{I+1}C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})-\frac{2}{I+1}C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal \\
&+\frac{1}{I+1}(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal](\boldsymbol{x}_0^F)^\intercal+\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal+\sigma^2-2\boldsymbol{x}_0^F[C+\frac{(\boldsymbol{I}_q-C)}{I+1}]\Sigma(\boldsymbol{x}_0^F)^\intercal
\end{aligned}$$

$=: \nu$.

Consequently,

$$\hat{Y}_0^F - Y_0^F \sim N(0, \nu). \tag{4.7}$$

□

If variance components are unknown the estimations of Σ and σ^2 are needed to construct an approximate prediction interval. In the following theorem, an approximate prediction interval for Y_0^F is obtained when $\hat{\Sigma}_{obs}$ and $\hat{\sigma}^2_{obs}$ are replaced in (4.7).

Theorem 4.1. *Approximate prediction interval for Y_0^F based on the modified Rao's prediction:*
Consider the defined RCR model with all assumptions for random effects and errors in Lemma 4.1 for all individuals including the new one. Let $\hat{Y}_0^F := \boldsymbol{x}_0^F(\hat{C}\hat{A}_{0,OLS} + (\boldsymbol{I}_q - \hat{C})\hat{\boldsymbol{\alpha}}_{obs})$, where $\hat{A}_{0,OLS} := (\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}Y_0^P$, be the prediction for the new observation at x_0^F. Then the variance of prediction error ν can be estimated as

$$\begin{aligned}
\hat{\nu} :=\ & \boldsymbol{x}_0^F[\hat{C}(\hat{\Sigma}_{obs}+\hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})\hat{C}^\intercal+\hat{v}-\hat{C}\hat{v}-\hat{v}\hat{C}^\intercal+\hat{C}\hat{v}\hat{C}^\intercal \\
&+\frac{1}{I+1}\hat{C}(\hat{\Sigma}_{obs}+\hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})-\frac{2}{I+1}\hat{C}(\hat{\Sigma}_{obs}+\hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})\hat{C}^\intercal \\
&+\frac{1}{I+1}(\hat{\Sigma}_{obs}+\hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})\hat{C}^\intercal](\boldsymbol{x}_0^F)^\intercal+\boldsymbol{x}_0^F\hat{\Sigma}_{obs}(\boldsymbol{x}_0^F)^\intercal+\hat{\sigma}^2_{obs} \\
&-2\boldsymbol{x}_0^F[\hat{C}+\frac{(\boldsymbol{I}_q-\hat{C})}{I+1}]\hat{\Sigma}_{obs}(\boldsymbol{x}_0^F)^\intercal,
\end{aligned}$$

where

$$\hat{C} := \hat{\Sigma}_{obs}[\hat{\Sigma}_{obs} + \hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}]^{-1},$$

$$\hat{v} = \frac{1}{(I+1)^2}[(I+1)\hat{\Sigma}_{obs} + \hat{\sigma}^2_{obs}(I(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1} + (\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})],$$

and $\hat{\Sigma}_{obs}$, $\hat{\sigma}^2_{obs}$ and $\hat{\boldsymbol{\alpha}}_{obs}$ have been defined in (4.3).
By use of $\hat{\nu}$, the standardized prediction error converges to the standard normal distribution for $N, N_0 \to \infty$. Consequently, the $(1-\alpha)$-prediction interval for Y_0^F is given by

$$\widehat{\mathcal{PI}}(\boldsymbol{y}_{obs}) = [\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}],$$

where $\hat{y}_0^F$ is the realization of $\hat{Y}_0^F$ and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of standard normal distribution.

Proof. Assume $B_1 := \lim_{N\to\infty}(N(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})$ and $B_2 := \lim_{N_0\to\infty}(N_0(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})$ exist and are positive definite. Hence, for large enough N_0 we approximately have

$$\hat{\Sigma}_{obs} + \hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1} \approx \hat{\Sigma}_{obs} \tag{4.8}$$

and

$$\Sigma + \sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1} \approx \Sigma. \tag{4.9}$$

Consequently,

$$\hat{C} \approx \hat{\Sigma}_{obs}\hat{\Sigma}^{-1}_{obs} = \boldsymbol{I}_q, \quad \text{and} \quad C \approx \Sigma\Sigma^{-1} = \boldsymbol{I}_q. \tag{4.10}$$

Now based on (4.8), (4.10) and the consistency of $\hat{\sigma}^2_{obs}$ and $\hat{\Sigma}_{obs}$, the asymptotic properties of $\hat{\nu}$ can be studied. If $N, N_0 \to \infty$ then

$$\text{plim}_{N,N_0\to\infty}(\hat{v}) = \text{plim}_{N,N_0\to\infty}\{\frac{1}{(I+1)^2}[(I+1)\hat{\Sigma}_{obs}+\hat{\sigma}^2_{obs}\{I(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}+(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\}]\}$$

$$= \frac{1}{(I+1)^2}[(I+1)\text{plim}_{N,N_0\to\infty}(\hat{\Sigma}_{obs})+\text{plim}_{N,N_0\to\infty}(\hat{\sigma}^2_{obs})\{I\lim_{N\to\infty}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}$$

$$+\lim_{N_0\to\infty}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\}],$$

$$= (\frac{1}{I+1})\frac{S_{obs}}{I}, \quad \text{(from (4.4))}$$

hence from (4.8) and (4.10),

$$\hat{\nu^*} := \text{plim}_{N,N_0\to\infty}(\hat{\nu})$$

$$= \text{plim}_{N,N_0\to\infty}\{\boldsymbol{x}_0^F[\hat{C}(\hat{\Sigma}_{obs}+\hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})\hat{C}^\intercal+\hat{v}-\hat{C}\hat{v}-\hat{v}\hat{C}^\intercal+\hat{C}\hat{v}\hat{C}^\intercal$$

$$+\frac{1}{I+1}\hat{C}(\hat{\Sigma}_{obs}+\hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})-\frac{2}{I+1}\hat{C}(\hat{\Sigma}_{obs}+\hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})\hat{C}^\intercal$$

$$+\frac{1}{I+1}(\hat{\Sigma}_{obs}+\hat{\sigma}^2_{obs}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})\hat{C}^\intercal](\boldsymbol{x}_0^F)^\intercal+\boldsymbol{x}_0^F\hat{\Sigma}_{obs}(\boldsymbol{x}_0^F)^\intercal+\hat{\sigma}^2_{obs}$$

$$-2\boldsymbol{x}_0^F[\hat{C}+\frac{(\boldsymbol{I}_q-\hat{C})}{I+1}]\hat{\Sigma}_{obs}(\boldsymbol{x}_0^F)^\intercal\}$$

$$= \boldsymbol{x}_0^F[\frac{S_{obs}}{I}+(\frac{1}{I+1})\frac{S_{obs}}{I}-(\frac{1}{I+1})\frac{S_{obs}}{I}-(\frac{1}{I+1})\frac{S_{obs}}{I}+(\frac{1}{I+1})\frac{S_{obs}}{I}+(\frac{1}{I+1})\frac{S_{obs}}{I}$$

$$-(\frac{2}{I+1})\frac{S_{obs}}{I}+(\frac{1}{I+1})\frac{S_{obs}}{I}](\boldsymbol{x}_0^F)^\intercal+\boldsymbol{x}_0^F[\frac{S_{obs}}{I}](\boldsymbol{x}_0^F)^\intercal+\sigma^2-2\boldsymbol{x}_0^F\frac{S_{obs}}{I}(\boldsymbol{x}_0^F)^\intercal$$

$= \sigma^2$.

Consequently, $\hat{\nu}$ is a consistent estimation for σ^2 but inconsistent for ν. For getting the asymptotic distribution of $\hat{Y}_0^F - Y_0^F$, study of the asymptotic properties of ν is required. By use of (4.9) we have

$$\lim_{N,N_0\to\infty}(v) = \lim_{N,N_0\to\infty}(\frac{1}{(I+1)^2}[(I+1)\Sigma+\sigma^2\{I(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}+(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\}])$$

$$= \frac{1}{(I+1)^2}[(I+1)\Sigma+\sigma^2\{I\lim_{N\to\infty}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})+\lim_{N_0\to\infty}(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\}]$$

$$= (\frac{1}{I+1})\Sigma,$$

therefore from (4.9) and (4.10),

$$\nu^* := \lim_{N,N_0\to\infty}(\nu) = \lim_{N,N_0\to\infty}[\boldsymbol{x}_0^F[C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal+v-Cv$$

$$-vC^\intercal+CvC^\intercal+\frac{1}{I+1}C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})-\frac{2}{I+1}C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal$$

$$+\frac{1}{I+1}(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal](\boldsymbol{x}_0^F)^\intercal+\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal+\sigma^2$$

$$-2\boldsymbol{x}_0^F[C+\frac{(\boldsymbol{I}_q-C)}{I+1}]\Sigma(\boldsymbol{x}_0^F)^\intercal]$$

$$= \boldsymbol{x}_0^F[\Sigma+(\frac{1}{I+1})\Sigma-(\frac{1}{I+1})\Sigma-(\frac{1}{I+1})\Sigma+(\frac{1}{I+1})\Sigma+\frac{1}{I+1}\Sigma-\frac{2}{I+1}\Sigma$$

$$+\frac{1}{I+1}\Sigma](\boldsymbol{x}_0^F)^\intercal+\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal+\sigma^2-2\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal$$

$$= \sigma^2.$$

Finally by applying the Slutzky theorem, the asymptotic distribution of prediction error is given by

$$\frac{\hat{Y}_0^F - Y_0^F}{\sqrt{\hat{\nu}}} \xrightarrow[N,N_0\to\infty]{D} \frac{Z}{\sqrt{\mathbb{V}\mathrm{ar}(Z)}} \sim \mathbb{N}(0,1),$$

where $Z \sim \mathbb{N}(0,\sigma^2)$. In conclusion, the approximate $(1-\alpha)$-prediction interval for Y_0^F from a partially observed individual is given by

$$[\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}]. \tag{4.11}$$

□

4.1.1.1 Extension of the method of Rao (eRao)

In this approach, we predict the future observation Y_0^F using the prediction procedure of Rao (see equation 4.1), in which the parameters are estimated based on the information of $\boldsymbol{Y}$. We extend this prediction procedure by adding an approximate prediction interval for Y_0^F in Corollary 4.1.

Corollary 4.1. The prediction interval for Y_0^F based on the Rao's prediction procedure can be obtained as a special case of the modified Rao's prediction interval and is given by

$$[\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}],$$

where $\hat{y}_0^F$ is the realization of $\hat{Y}_0^F$ and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of standard normal distribution and

$$\hat{\nu} := \boldsymbol{x}_0^F[\hat{C}(\hat{\Sigma}+\hat{\sigma}^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})\hat{C}^\intercal+\hat{v}-\hat{C}\hat{v}-\hat{v}\hat{C}^\intercal+\hat{C}\hat{v}\hat{C}^\intercal](\boldsymbol{x}_0^F)^\intercal+\boldsymbol{x}_0^F\hat{\Sigma}(\boldsymbol{x}_0^F)^\intercal+\hat{\sigma}^2$$
$$-2\boldsymbol{x}_0^F\hat{C}\hat{\Sigma}(\boldsymbol{x}_0^F)^\intercal,$$

where

$$\hat{C} := \hat{\Sigma}[\hat{\Sigma} + \hat{\sigma}^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}]^{-1}, \quad \hat{v} = \frac{1}{I}(\hat{\Sigma} + \hat{\sigma}^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}),$$

and $\hat{\Sigma}$ and $\hat{\sigma}^2$ have been defined respectively in (3.7) and (3.6).

Proof. Consider $v := \mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{GLS}) = \frac{1}{I}(\Sigma + \sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})$, where $\hat{\boldsymbol{\alpha}}_{GLS}$ has been defined in Theorem 3.1 and $C = \Sigma[\Sigma + \sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}]^{-1}$. Because of the independency between $\hat{A}_{0,OLS}$ and $\hat{\boldsymbol{\alpha}}_{GLS}$, the expression of $\mathbb{V}\text{ar}(\hat{Y}_0^F)$ in the proof of Lemma 4.1 reduces to the following expression:

$$\mathbb{V}\text{ar}(\hat{Y}_0^F) = \boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS})C^\intercal+(\boldsymbol{I}_q-C)\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{GLS})(\boldsymbol{I}_q-C)^\intercal](\boldsymbol{x}_0^F)^\intercal$$
$$= \boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS})C^\intercal+\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{GLS})-C\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{GLS})$$
$$-\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{GLS})C^\intercal+C\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{GLS})C^\intercal](\boldsymbol{x}_0^F)^\intercal$$
$$= \boldsymbol{x}_0^F[C(\Sigma+\sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal+v-Cv-vC^\intercal+CvC^\intercal](\boldsymbol{x}_0^F)^\intercal.$$

Now by use of the independencies between $\hat{\boldsymbol{\alpha}}_{GLS}$ and A_0, between $\hat{\boldsymbol{\alpha}}_{GLS}$ and E_0^F and between $\hat{A}_{0,OLS}$ and E_0^F, the expression of $\mathbb{C}\text{ov}(\hat{Y}_0^F, Y_0^F)$ in the proof of Lemma 4.1 reduces to the following expression:

$$\mathbb{C}\text{ov}(\hat{Y}_0^F, Y_0^F) = \mathbb{C}\text{ov}(\boldsymbol{x}_0^F[C\hat{A}_{0,OLS}+(\boldsymbol{I}_q-C)\hat{\boldsymbol{\alpha}}_{GLS}], \boldsymbol{x}_0^F A_0+E_0^F)$$
$$= \mathbb{C}\text{ov}[\boldsymbol{x}_0^F(C\hat{A}_{0,OLS}+(\boldsymbol{I}_q-C)\hat{\boldsymbol{\alpha}}_{GLS}), \boldsymbol{x}_0^F A_0]$$

$+\mathbb{C}\text{ov}[\boldsymbol{x}_0^F(C\hat{A}_{0,OLS}+(\boldsymbol{I}_q-C)\hat{\boldsymbol{\alpha}}_{GLS}), E_0^F]$

$= \boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS}, A_0)+(\boldsymbol{I}_q-C)\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{GLS}, A_0)](\boldsymbol{x}_0^F)^\intercal$

$+\boldsymbol{x}_0^F[C\mathbb{C}\text{ov}(\hat{A}_{0,OLS}, E_0^F)+(\boldsymbol{I}_q-C)\mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{GLS}, E_0^F)]$

$= \boldsymbol{x}_0^F C\mathbb{C}\text{ov}(\hat{A}_{0,OLS}, A_0)(\boldsymbol{x}_0^F)^\intercal$

$= \boldsymbol{x}_0^F C\Sigma(\boldsymbol{x}_0^F)^\intercal.$

Finally, the variance of prediction error is given by

$$\mathbb{V}\text{ar}(\hat{Y}_0^F - Y_0^F) := \nu = \boldsymbol{x}_0^F[C(\Sigma + \sigma^2(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1})C^\intercal + v - Cv - vC^\intercal + CvC^\intercal](\boldsymbol{x}_0^F)^\intercal$$

$$+\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal+\sigma^2-2\boldsymbol{x}_0^F C\Sigma(\boldsymbol{x}_0^F)^\intercal.$$

The estimate of ν ($\hat{\nu}$) can be obtained by replacing the estimates of Σ and σ^2. Now by considering this special case in Theorem 4.1, the approximate prediction interval for Y_0^F is given by:

$$\widehat{\mathcal{PI}}(\boldsymbol{y}_{obs}) = [\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}].$$

□

4.1.2 The modified method of Swamy (mSwamy): The convex combination of Swamy's prediction approach and classical approach for linear models

Let the defined RCR model (3.1) with the assumptions [A1]-[A5] for the balanced data. Consider $\hat{\boldsymbol{\alpha}}_{GLS}$, $\hat{\Sigma}$ and $\hat{\sigma}^2$ be obtained based on the information of $\boldsymbol{Y} := (Y_1, \cdots, Y_I)$ such that $\hat{\boldsymbol{\alpha}}_{GLS} = \frac{1}{I}\sum_{i=1}^I \hat{A}_{i,OLS}$, where $\hat{A}_{i,OLS} := (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal Y_i$, and $\hat{\Sigma}$ and $\hat{\sigma}^2$ are obtained from (3.6) and (3.7). From the Swamy's prediction, the predictor of the future observation at $\boldsymbol{x}_0^F$ is given by

$$\hat{Y}_0^{F(old)} = \boldsymbol{x}_0^F\hat{\boldsymbol{\alpha}}_{GLS}.$$

Based on Lemma 3.2, the prediction error $(\hat{Y}_0^{F(old)} - Y_0^F)$ has normal distribution with mean 0 and variance ν

$$\nu := \frac{1}{I}\boldsymbol{x}_0^F(\Sigma + \sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal + \boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + \sigma^2.$$

If the future observation at x_0^F is predicted only based on the information of Y_0^P, then the predictor is given by

$$\hat{Y}_0^{F(new)} = \boldsymbol{x}_0^F\hat{A}_{0,OLS},$$

where $\hat{A}_{0,OLS} := (\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}Y_0^P$.
Our proposed combined predictor based on the information of $\boldsymbol{Y}_{obs}$ is given by

$$\hat{Y}_0^F = (1-w)\hat{Y}_0^{F(old)} + w\hat{Y}_0^{F(new)},$$

where $w = \frac{k}{k+h}$, with $k := \frac{1}{I}\boldsymbol{x}_0^F(\Sigma + \sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal + \boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal$, and
$h := \sigma^2\boldsymbol{x}_0^F(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}(\boldsymbol{x}_0^F)^\intercal$, which minimizes the variance of prediction error.

Lemma 4.2. *Prediction interval for Y_0^F based on the modified method of Swamy: Consider the RCR model (3.1) and the assumptions [A1]-[A5] for the balanced data. Let the observations of the new individual be modeled as in Lemma 4.1. The prediction error $(\hat{Y}_0^F - Y_0^F)$ has normal distribution with mean 0 and variance v as follows:*

$$v := (1-w)^2k + w^2h + \sigma^2,$$

where

$$k := \frac{1}{I}\boldsymbol{x}_0^F(\Sigma + \sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal + \boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal,$$

$$h := \sigma^2\boldsymbol{x}_0^F(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}(\boldsymbol{x}_0^F)^\intercal, \quad w = \frac{k}{k+h}.$$

Then the $(1-\alpha)$-prediction interval is given by

$$[\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}], \tag{4.12}$$

where $\hat{y}_0^F$ is the realization of $\hat{Y}_0^F$ and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of standard normal distribution and $\hat{\nu}$ is obtained by replacing the estimates $\hat{\Sigma}$ and $\hat{\sigma}^2$ from (3.6) and (3.7) in w, k and h.

Proof. Because

$$\mathbb{E}(\hat{Y}_0^{F(new)}) = \mathbb{E}(\hat{Y}_0^{F(old)}) = \boldsymbol{x}_0^F\boldsymbol{\alpha},$$

therefore, $\mathbb{E}(\hat{Y}_0^F - Y_0^F) = 0$. The term $w\hat{Y}_0^{F(new)} - Y_0^F$ can be simplified as

$$w\hat{Y}_0^{F(new)} - Y_0^F = w\boldsymbol{x}_0^F(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}(\boldsymbol{X}_0^P A_0 + E_0^P) - \boldsymbol{x}_0^F A_0 - E_0^F$$

$$= w\boldsymbol{x}_0^F A_0 + w\boldsymbol{x}_0^F(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}E_0^P - \boldsymbol{x}_0^F A_0 - E_0^F$$

$$= (w-1)\boldsymbol{x}_0^F A_0 + w\boldsymbol{x}_0^F(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}E_0^P - E_0^F.$$

Hence, by use of the independency between A_0, E_0^P and E_0^F the variance of $w\hat{Y}_0^{F(new)} - Y_0^F$ is given by

$$\mathbb{V}\mathrm{ar}(w\hat{Y}_0^{F(new)} - Y_0^F) = (1-w)^2\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + w^2\sigma^2\boldsymbol{x}_0^F(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}(\boldsymbol{x}_0^F)^\intercal + \sigma^2.$$

By the fact that

$$\mathbb{V}\mathrm{ar}((1-w)\hat{Y}_0^{F(old)}) = (1-w)^2 \cdot \frac{1}{I}\boldsymbol{x}_0^F(\Sigma + \sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal,$$

and based on the independency between $\hat{Y}_0^{F(old)}$ and $w\hat{Y}_0^{F(new)} - Y_0^F$ we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}((1-w)\hat{Y}_0^{F(old)} + w\hat{Y}_0^{F(new)} - Y_0^F) &= (1-w)^2k + w^2h + \sigma^2 \\ &=: v.\end{aligned}$$

By replacing the estimates $\hat{\Sigma}$ and $\hat{\sigma}^2$ from (3.6) and (3.7) in w, k and h, the $(1-\alpha)$-prediction interval for Y_0^F is obtained as follows:

$$[\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}].$$

□

4.1.3 Approach of Liski and Nummi in the unbalanced case

Liski and Nummi (1996) suggest a general estimation procedure using the EM algorithm by considering some unobserved response values as missing part of the data set. In the following, we present their suggested EM algorithm.

4.1.3.1 Estimation method of Liski and Nummi (1996)

Consider the following RCR model

$$Y_i = \boldsymbol{X}_i A_i + E_i, \quad i \in \{1, \cdots, I\} \tag{4.13}$$

where $Y_i := (Y_{i1}, \cdots, Y_{iN_i})^\intercal$ is the observation vector of the ith individual and $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}^\intercal, \cdots, \boldsymbol{x}_{iN_i}^\intercal)^\intercal$ is the $N_i \times q$ covariate matrix of the ith individual with $\boldsymbol{x}_{ij} = (1, h_1(x_{ij}), \cdots, h_{q-1}(x_{ij}))$. Suppose $A_i \overset{iid}{\sim} \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$ and $E_i \overset{iid}{\sim} \mathbb{N}_{N_i}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{N_i})$. Consider the presented RCR model for observations of a new individual with the proposed partitions in Lemma 4.1. Suppose that $A_0 \sim \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$, $E_0 \sim \mathbb{N}_{N_0+1}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{N_0+1})$ and $E_0^P, E_0^F, E_1, \cdots, E_I, A_0, A_1, \cdots, A_I$ are independent.

The E step:
Let $\boldsymbol{Y}_{obs} = (Y_1^\intercal, \cdots, Y_I^\intercal, Y_0^{P\intercal})^\intercal$ and $\boldsymbol{Z} = (\boldsymbol{Y}_{obs}, Y_0^F)$ represent respectively the observed and complete data such that Y_0^F represents the missing data and also $\boldsymbol{y}_{obs}$ and $\boldsymbol{z}$ be the realizations of $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Z}$. For simplifying calculations, apply the

positive definite matrix $D \in \mathbb{R}^{q\times q}$ such that $\Sigma = \sigma^2 D$ and $\boldsymbol{\theta} := (\theta_1, \cdots, \theta_s)$ such that $D = D(\boldsymbol{\theta})$. Now for $i = 1, \cdots, I$, we have $\mathbb{E}(Y_i) = \boldsymbol{X}_i\boldsymbol{\alpha}$ and

$$\mathbb{C}\text{ov}(Y_i) = \boldsymbol{X}_i\Sigma\boldsymbol{X}_i^\intercal + \sigma^2\boldsymbol{I}_{N_i} = \boldsymbol{X}_i(\sigma^2 D(\boldsymbol{\theta}))\boldsymbol{X}_i^\intercal + \sigma^2\boldsymbol{I}_{N_i}$$

$$= \sigma^2[\boldsymbol{X}_i D(\boldsymbol{\theta})\boldsymbol{X}_i^\intercal + \boldsymbol{I}_{N_i}] =: \sigma^2 H_i(\boldsymbol{\theta}).$$

And for the new individual we have similarly $\mathbb{E}(Y_0) = \boldsymbol{X}_0\boldsymbol{\alpha}$ and

$$\mathbb{C}\text{ov}(Y_0) = \sigma^2[\boldsymbol{X}_0 D(\boldsymbol{\theta})\boldsymbol{X}_0^\intercal + \boldsymbol{I}_{N_0+1}] =: \sigma^2 H_0(\boldsymbol{\theta}).$$

By considering $\boldsymbol{\beta} := (\boldsymbol{\alpha}, \boldsymbol{\theta}, \sigma^2)$, we define $Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)})$ as follows:

$$Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}) = \mathbb{E}(\log f(\boldsymbol{Z}|\boldsymbol{\beta})|\boldsymbol{y}_{obs}, \boldsymbol{\beta}^{(k)}) = \int (\log f(\boldsymbol{z}|\boldsymbol{\beta}))f(y_0^F|y_0^P, \boldsymbol{\beta}^{(k)})dy_0^F,$$

where

$$\log f(\boldsymbol{z}|\boldsymbol{\beta}) = -\frac{1}{2}[(N+N_0+1)\log(2\pi\sigma^2) + \sum_{i=0}^{I}\log|H_i(\boldsymbol{\theta})|$$

$$+\frac{\sum_{i=0}^{I}(y_i - \boldsymbol{X}_i\boldsymbol{\alpha})^\intercal H_i^{-1}(\boldsymbol{\theta})(y_i - \boldsymbol{X}_i\boldsymbol{\alpha})}{\sigma^2}],$$

with $N := \sum_{i=1}^{I} N_i$, and $y_0 = (y_0^{P\intercal}, y_0^F)^\intercal$ and

$$H_i(\boldsymbol{\theta}) := \boldsymbol{X}_i D(\boldsymbol{\theta})\boldsymbol{X}_i^\intercal + \boldsymbol{I}_{N_i}, \quad i \in \{1, \cdots, I\}, \quad H_0(\boldsymbol{\theta}) := \begin{pmatrix} H_0^P(\boldsymbol{\theta}) & H_0^{FP\intercal}(\boldsymbol{\theta}) \\ H_0^{FP}(\boldsymbol{\theta}) & H_0^F(\boldsymbol{\theta}) \end{pmatrix},$$

where y_i, y_0^P and y_0^F are respectively the realizations of Y_i, Y_0^P and Y_0^F and

$$H_0^P(\boldsymbol{\theta}) := \boldsymbol{X}_0^P D(\boldsymbol{\theta})\boldsymbol{X}_0^{P\intercal} + \boldsymbol{I}_{N_0}, \quad H_0^F(\boldsymbol{\theta}) := \boldsymbol{x}_0^F D(\boldsymbol{\theta})\boldsymbol{x}_0^{F\intercal} + 1,$$

and

$$H_0^{FP}(\boldsymbol{\theta}) := \boldsymbol{x}_0^F D(\boldsymbol{\theta})\boldsymbol{X}_0^{P\intercal}.$$

The M step:

Let $\hat{\boldsymbol{\beta}}$ be the unique solution at the $(k+1)$th iteration for the following equation:

$$\frac{\partial}{\partial\boldsymbol{\beta}}Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0.$$

By allowing differentiation under the integral sign (Jamshidian and Jennrich, 1993) we have

$$\mathbb{E}\{\frac{\partial}{\partial \boldsymbol{\beta}} \log f(\boldsymbol{Z}|\boldsymbol{\beta})|\boldsymbol{y}_{obs}, \boldsymbol{\beta}^{(k)}\}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0. \tag{4.14}$$

Then

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \log f(\boldsymbol{z}|\boldsymbol{\beta}) = \frac{1}{\sigma^2}[\sum_{i=0}^{I} y_i^\intercal H_i^{-1}(\boldsymbol{\theta})\boldsymbol{X}_i - \boldsymbol{\alpha}^\intercal \boldsymbol{X}_i^\intercal H_i^{-1}(\boldsymbol{\theta})\boldsymbol{X}_i],$$

moreover,

$$\frac{\partial}{\partial \sigma^2} \log f(\boldsymbol{z}|\boldsymbol{\beta}) = -\frac{N+N_0+1}{2\sigma^2} + \frac{\sum_{i=0}^{I}(y_i - \boldsymbol{X}_i\boldsymbol{\alpha})^\intercal H_i^{-1}(\boldsymbol{\theta})(y_i - \boldsymbol{X}_i\boldsymbol{\alpha})}{2\sigma^4}$$

and for $w = 1, \cdots, s$

$$\frac{\partial}{\partial \theta_w} \log f(\boldsymbol{z}|\boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=0}^{I} \mathrm{tr}(H_i^{-1}(\boldsymbol{\theta})\frac{\partial H_i(\boldsymbol{\theta})}{\partial \theta_w})$$

$$+\frac{1}{2\sigma^2}\sum_{i=0}^{I}(y_i - \boldsymbol{X}_i\boldsymbol{\alpha})^\intercal H_i^{-1}(\boldsymbol{\theta})\frac{\partial H_i(\boldsymbol{\theta})}{\partial \theta_w} H_i^{-1}(\boldsymbol{\theta})(y_i - \boldsymbol{X}_i\boldsymbol{\alpha}).$$

According to the above expressions, the unique solution of (4.14) at the $(k+1)$th iteration is given by

$$\hat{\boldsymbol{\alpha}}^{(k+1)} = (\sum_{i=0}^{I} \boldsymbol{X}_i^\intercal H_i^{-1}(\boldsymbol{\theta}^{(k)})\boldsymbol{X}_i)^{-1}[\sum_{i=1}^{I}(\boldsymbol{X}_i^\intercal H_i^{-1}(\boldsymbol{\theta}^{(k)})y_i)$$

$$+\boldsymbol{X}_0^\intercal H_0^{-1}(\boldsymbol{\theta}^{(k)})\mathbb{E}\{\begin{pmatrix} Y_0^P \\ Y_0^F \end{pmatrix}|y_0^P, \boldsymbol{\beta}^{(k)}\}], \tag{4.15}$$

$$\hat{\sigma}^{2(k+1)} = \frac{1}{N+N_0+1}[\sum_{i=1}^{I}(y_i - \boldsymbol{X}_i\boldsymbol{\alpha}^{(k)})^\intercal H_i^{-1}(\boldsymbol{\theta}^{(k)})(y_i - \boldsymbol{X}_i\boldsymbol{\alpha}^{(k)}),$$

$$+\mathbb{E}\{\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}^\intercal H_0^{-1}(\boldsymbol{\theta}) \begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}|y_0^P, \boldsymbol{\beta}^{(k)}\}], \tag{4.16}$$

and

$$\frac{1}{\sigma^{2(k)}}[\sum_{i=1}^{I}(y_i - \boldsymbol{X}_i\boldsymbol{\alpha}^{(k)})^\intercal H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial \theta_w} H_i^{-1}(\boldsymbol{\theta}^{(k)})(y_i - \boldsymbol{X}_i\boldsymbol{\alpha}^{(k)})$$

$$+\mathbb{E}\{\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}^\intercal H_0^{-1}(\boldsymbol{\theta})\frac{\partial H_0(\boldsymbol{\theta})}{\partial \theta_w}H_0^{-1}(\boldsymbol{\theta})\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}|y_0^P, \boldsymbol{\beta}^{(k)}\}]$$

$$= \sum_{i=0}^{I}\mathrm{tr}(H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial \theta_w}). \tag{4.17}$$

The calculation of conditional expectations has been presented in the appendix.

Estimation of Standard errors [Demidenko (2004)]
Demidenko (2004) presents some formulas for the asymptotic covariance matrix of maximum likelihood estimates in linear mixed effects models. If $\hat{\boldsymbol{\alpha}}$, $\hat{\Sigma}$ and $\hat{\sigma}^2$ are the ML estimates at convergence then the asymptotic covariance matrix of $\hat{\boldsymbol{\alpha}}$ is approximately given by

$$\mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\alpha}}) = [\sum_{i=1}^{I}\boldsymbol{X}_i^\intercal V_i^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{X}_i + \boldsymbol{X}_0^\intercal V_0(\hat{\boldsymbol{\theta}})\boldsymbol{X}_0]^{-1}, \tag{4.18}$$

where

$$V_i(\hat{\boldsymbol{\theta}}) := \boldsymbol{X}_i\hat{\Sigma}\boldsymbol{X}_i^\intercal + \hat{\sigma}^2\boldsymbol{I}_{N_i}, \quad V_0(\hat{\boldsymbol{\theta}}) := \begin{pmatrix} V_0^P(\hat{\boldsymbol{\theta}}) & V_0^{FP\intercal}(\hat{\boldsymbol{\theta}}) \\ V_0^{FP}(\hat{\boldsymbol{\theta}}) & V_0^F(\hat{\boldsymbol{\theta}}) \end{pmatrix},$$

with

$$V_0^P(\hat{\boldsymbol{\theta}}) := \boldsymbol{X}_0^P\hat{\Sigma}\boldsymbol{X}_0^{P\intercal} + \hat{\sigma}^2\boldsymbol{I}_{N_0}, \quad V_0^{FP}(\hat{\boldsymbol{\theta}}) := \boldsymbol{x}_0^F\hat{\Sigma}\boldsymbol{X}_0^{P\intercal},$$
$$V_0^F(\hat{\boldsymbol{\theta}}) := \boldsymbol{x}_0^F\hat{\Sigma}\boldsymbol{x}_0^{F\intercal} + \hat{\sigma}^2.$$

Consider the scaled covariance matrix of random effects as $\Sigma = \sigma^2 D$. By use of D, Demidenko (2004) simplifies the log likelihood function in linear mixed effects models and presents the information matrix for the variance components. Consider $\boldsymbol{\theta} := (\sigma^2, \mathrm{vech}^\intercal(D))^\intercal$ and $\hat{\boldsymbol{\theta}} := (\hat{\sigma}^2, \mathrm{vech}^\intercal(\hat{D}))^\intercal$, where $\mathrm{vech}(D) := (D_{11}, D_{21}, \cdots, D_{q1}, D_{22}, \cdots, D_{qq})^\intercal$, then the asymptotic variance covariance matrix of the ML estimate $\hat{\boldsymbol{\theta}}$ is given by

$$\mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\theta}}) = 2\begin{pmatrix} (\sum_{i=1}^{I}N_i + (N_0+1))\sigma^{-4} & \sigma^{-2}\mathrm{vec}^\intercal(\sum_{i=0}^{I}R_i)\mathfrak{D}^{+\intercal} \\ \sigma^{-2}\mathfrak{D}^+\mathrm{vec}(\sum_{i=0}^{I}R_i) & \mathfrak{D}^+(\sum_{i=0}^{I}R_i \otimes R_i)\mathfrak{D}^{+\intercal} \end{pmatrix}^{-1}, \tag{4.19}$$

with $R_i := \boldsymbol{X}_i^\intercal(\boldsymbol{I}_{N_i} + \boldsymbol{X}_i D\boldsymbol{X}_i^\intercal)^{-1}\boldsymbol{X}_i$ for $i = 1, \cdots, I$, $R_0 := \boldsymbol{X}_0^\intercal(\boldsymbol{I}_{N_0+1} + \boldsymbol{X}_0 D\boldsymbol{X}_0^\intercal)^{-1}\boldsymbol{X}_0$ and $\mathfrak{D}^+ = (\mathfrak{D}_q^\intercal\mathfrak{D}_q)^{-1}\mathfrak{D}_q^\intercal$, where $\mathfrak{D}_q$ is the $q^2 \times q(q+1)/2$ duplication matrix (see Appendix). By applying matrix block inverse (see Appendix), we have

$$\mathbb{V}\mathrm{ar}(\hat{\sigma}^2) = \frac{2\sigma^4}{\sum_{i=1}^{I}N_i + (N_0+1) - w}, \tag{4.20}$$

and

$$\mathbb{C}\text{ovh}(\hat{D}) := \mathbb{C}\text{ov}(\text{vech}(\hat{D}))$$

$$= 2[\mathfrak{D}^{+}(\sum_{i=0}^{I} R_i \otimes R_i - \frac{1}{\sum_{i=1}^{I} N_i + (N_0+1)} \text{vec}(\sum_{i=0}^{I} R_i)\text{vec}^{\intercal}(\sum_{i=0}^{I} R_i))\mathfrak{D}^{+\intercal}]^{-1}, \quad (4.21)$$

with

$$w := \text{vec}^{\intercal}(\sum_{i=0}^{I} R_i)\mathfrak{D}^{+\intercal}[\mathfrak{D}^{+}(\sum_{i=0}^{I} R_i \otimes R_i)\mathfrak{D}^{+\intercal}]^{-1}\mathfrak{D}^{+}\text{vec}(\sum_{i=0}^{I} R_i).$$

The variance covariance matrices can be estimated by replacing the maximum likelihood estimates $\hat{\boldsymbol{\alpha}}$, $\hat{D}$ and $\hat{\sigma}^2$ in (4.19)-(4.21). Since $\boldsymbol{\theta} := (\sigma^2, \text{vech}(D))$ and $\hat{\boldsymbol{\theta}} := (\hat{\sigma}^2, \text{vech}(\hat{D}))$, hence we can define $K(\boldsymbol{\theta}) := \sigma^2\text{vech(D)}$ and $K(\hat{\boldsymbol{\theta}}) := \hat{\sigma}^2\text{vech}(\hat{D})$. Let $\text{vech}(D)$ be an $m \times 1$ vector. Now by knowing that $\text{vech}(\hat{\Sigma}) = \hat{\sigma}^2\text{vech}(\hat{D})$, and based on the delta method we have

$$\mathbb{C}\text{ovh}(\hat{\Sigma}) := \mathbb{C}\text{ov}(\text{vech}(\hat{\Sigma})) = \mathbb{C}\text{ov}(\hat{\sigma}^2\text{vech}(\hat{D})) = \mathbb{C}\text{ov}(K(\hat{\boldsymbol{\theta}}))$$

$$\approx (\frac{\partial K(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}})\mathbb{C}\text{ov}(\hat{\boldsymbol{\theta}})(\frac{\partial K(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}})^{\intercal} = (\frac{\partial K(\boldsymbol{\theta})}{\partial \sigma^2}, \frac{\partial K(\boldsymbol{\theta})}{\partial \text{vech}(D)})\mathbb{C}\text{ov}(\hat{\boldsymbol{\theta}}) \begin{pmatrix} (\frac{\partial K(\boldsymbol{\theta})}{\partial \sigma^2})^{\intercal} \\ (\frac{\partial K(\boldsymbol{\theta})}{\partial \text{vech}(D)})^{\intercal} \end{pmatrix}$$

$$= (\text{vech}(D), \sigma^2 \boldsymbol{I}_m) \begin{pmatrix} \mathbb{V}\text{ar}(\hat{\sigma}^2) & \mathbb{C}\text{ov}(\hat{\sigma}^2, \text{vech}(\hat{D})) \\ \mathbb{C}\text{ov}^{\intercal}(\hat{\sigma}^2, \text{vech}(\hat{D})) & \mathbb{C}\text{ov}(\text{vech}(\hat{D})) \end{pmatrix} \begin{pmatrix} \text{vech}^{\intercal}(D) \\ \sigma^2 \boldsymbol{I}_m \end{pmatrix}$$

$$= [\text{vech}(D)\mathbb{V}\text{ar}(\hat{\sigma}^2) + \sigma^2\mathbb{C}\text{ov}^{\intercal}(\hat{\sigma}^2, \text{vech}(\hat{D}))]\text{vech}^{\intercal}(D)$$

$$+\sigma^2[\text{vech}(D)\mathbb{C}\text{ov}(\hat{\sigma}^2, \text{vech}(\hat{D})) + \sigma^2\mathbb{C}\text{ov}(\text{vech}(\hat{D}))]$$

$$= \mathbb{V}\text{ar}(\hat{\sigma}^2)\text{vech}(D)\text{vech}^{\intercal}(D) + \sigma^4\mathbb{C}\text{ovh}(\hat{D})$$

$$+\sigma^2[\text{vech}(D)\mathbb{C}\text{ov}(\hat{\sigma}^2, \text{vech}(\hat{D})) + \mathbb{C}\text{ov}^{\intercal}(\hat{\sigma}^2, \text{vech}(\hat{D}))\text{vech}^{\intercal}(D)], \quad (4.22)$$

where $\mathbb{C}\text{ov}(\hat{\sigma}^2, \text{vech}(\hat{D}))$ is the (2,1)th block of the matrix (4.19) (Demidenko, 2004).

4.1.3.2 Prediction Interval

Consider the RCR model (4.13) and the presented RCR model for the observations of a new individual with the proposed decompositions in Lemma 4.1. Suppose that the missing observation from the partially observed individual Y_0^F is modeled as $Y_0^F = \boldsymbol{x}_0^F A_0 + E_0^F$ and satisfies in the RCR model assumptions. Let $\boldsymbol{\theta} :=$

$(\sigma^2, \text{vech}^\intercal(D))^\intercal$ and $\hat{\boldsymbol{\theta}} := (\hat{\sigma}^2, \text{vech}^\intercal(\hat{D}))^\intercal$, such that $\Sigma = \sigma^2 D$ and $\hat{\Sigma} = \hat{\sigma}^2 \hat{D}$. Liski and Nummi (1996) propose $\hat{y}_0^F$ as the prediction of Y_0^F which is given by

$$\hat{y}_0^F = \boldsymbol{x}_0^F \hat{\boldsymbol{\alpha}} + V_0^{FP}(\hat{\boldsymbol{\theta}}) V_0^{P^{-1}}(\hat{\boldsymbol{\theta}})(y_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}}), \tag{4.23}$$

with

$$V_0^{FP}(\hat{\boldsymbol{\theta}}) := \boldsymbol{x}_0^F \hat{\Sigma} \boldsymbol{X}_0^{P\intercal}, \quad V_0^P(\hat{\boldsymbol{\theta}}) := \boldsymbol{X}_0^P \hat{\Sigma} \boldsymbol{X}_0^{P\intercal} + \hat{\sigma}^2 \boldsymbol{I}_{N_0},$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\theta}}$ are the obtained ML estimates from EM algorithm. In the following lemma, the prediction interval proposed by Liski and Nummi (1996) is presented and a proof for that is proposed.

Lemma 4.3. *Prediction interval for Y_0^F based on the Liski and Nummi's prediction [Liski and Nummi (1996)]:*
Consider the RCR model (4.13) and the presented RCR model for the observations of a new individual with the proposed decompositions in Lemma 4.1. Assume that the missing observation from the partially observed individual is modeled as $Y_0^F = \boldsymbol{x}_0^F A_0 + E_0^F$ and satisfies in the RCR model assumptions. If the variance components vector $\boldsymbol{\theta} := (\sigma^2, vech^\intercal(D))^\intercal$ is known (where $vech(D) := (D_{11}, D_{21}, \cdots, D_{q1}, D_{22}, \cdots, D_{qq})^\intercal$), then the prediction of Y_0^F is given by

$$\tilde{Y}_0^F = \boldsymbol{x}_0^F \tilde{\boldsymbol{\alpha}} + V_0^{FP} V_0^{P^{-1}}(Y_0^P - \boldsymbol{X}_0^P \tilde{\boldsymbol{\alpha}}), \tag{4.24}$$

with

$$\tilde{\boldsymbol{\alpha}} = [\sum_{i=1}^{I} \boldsymbol{X}_i^\intercal V_i^{-1} \boldsymbol{X}_i + \boldsymbol{X}_0^{P\intercal} V_0^{P^{-1}} \boldsymbol{X}_0^P]^{-1} (\sum_{i=1}^{I} \boldsymbol{X}_i^\intercal V_i^{-1} Y_i + \boldsymbol{X}_0^{P\intercal} V_0^{P^{-1}} Y_0^P),$$

$$\tilde{\boldsymbol{\alpha}} \sim \mathbb{N}_q(\boldsymbol{\alpha}, [\sum_{i=1}^{I} \boldsymbol{X}_i^\intercal V_i^{-1} \boldsymbol{X}_i + \boldsymbol{X}_0^P V_0^{P^{-1}} \boldsymbol{X}_0^P]^{-1}), \tag{4.25}$$

where

$$V_i := \mathbb{C}ov(Y_i) = \boldsymbol{X}_i \Sigma \boldsymbol{X}_i^\intercal + \sigma^2 \boldsymbol{I}_{N_i}, \quad i \in \{1, \cdots, I\}$$
$$V_0^P := \mathbb{C}ov(Y_0^P) = \boldsymbol{X}_0^P \Sigma \boldsymbol{X}_0^{P\intercal} + \sigma^2 \boldsymbol{I}_{N_0}.$$

The $(1-\alpha)$-prediction interval for Y_0^F is given by

$$\mathcal{PI}(\boldsymbol{y}_{obs}) = [\tilde{y}_0^F - q_{1-\alpha/2}\sqrt{\Omega_0}, \tilde{y}_0^F + q_{1-\alpha/2}\sqrt{\Omega_0}],$$

where $\tilde{y}_0^F$ is the realization of $\tilde{Y}_0^F$ and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of standard normal distribution,

$$\Omega_0 := \mathbb{V}ar(\tilde{Y}_0^F - Y_0^F) = \Omega_0^F + M_0[\mathbb{C}ov(\tilde{\boldsymbol{\alpha}})]M_0^\intercal,$$

$$\begin{aligned}\Omega_0^F &= \mathbb{V}ar(Y_0^F) - \mathbb{C}ov(Y_0^F, Y_0^P)[\mathbb{C}ov(Y_0^P)]^{-1}\mathbb{C}ov(Y_0^P, Y_0^F) \\ &= V_0^F - V_0^{FP}(V_0^P)^{-1}V_0^{FP\intercal},\end{aligned}$$

and

$$M_0 = \boldsymbol{x}_0^F - \mathbb{C}ov(Y_0^F, Y_0^P)[\mathbb{C}ov(Y_0^P)]^{-1}\boldsymbol{X}_0^P = \boldsymbol{x}_0^F - V_0^{FP}V_0^{P^{-1}}\boldsymbol{X}_0^P,$$

with

$$V_0^F := \mathbb{V}ar(Y_0^F) = \boldsymbol{x}_0^F\Sigma\boldsymbol{x}_0^{F\intercal} + \sigma^2, \quad V_0^{FP} := \mathbb{C}ov(Y_0^F, Y_0^P) = \boldsymbol{x}_0^F\Sigma\boldsymbol{X}_0^{P\intercal}.$$

Proof. The proof can be found in the appendix. □

Practically, the variance components vector $\boldsymbol{\theta}$ is unknown hence the ML estimate of $\boldsymbol{\theta}$ is needed. By substituting the obtained ML estimate $\hat{\boldsymbol{\theta}} := (\hat{\sigma}^2, \text{vech}^\intercal(\hat{D}))$ from the Liski and Nummi's EM algorithm, the variance of prediction error is estimated.

4.2 Prediction of the future observation in nonlinear random effects models

4.2.1 Combination of methods of Pinheiro/Bates and Liski/Nummi in the unbalanced case (PBLiski)

Consider the following nonlinear random effects model

$$Y_i = g_i(A_i, \boldsymbol{x}_i) + E_i, \quad i \in \{1, \cdots, I\} \tag{4.26}$$

with

$$g_i(A_i, \boldsymbol{x}_i) : \mathbb{R}^q \times \mathbb{R}^{N_i} \to \mathbb{R}^{N_i}, \quad g_i(A_i, \boldsymbol{x}_i) = \begin{pmatrix} g(A_i, x_{i1}) \\ \vdots \\ g(A_i, x_{iN_i}) \end{pmatrix}$$

and

$$A_i = \boldsymbol{\alpha} + \boldsymbol{\delta}_i, \quad \boldsymbol{\delta}_i \overset{iid}{\sim} \mathbb{N}_q(\mathbf{0}, \Sigma),$$

where Y_i is the N_i-dimensional response vector for the ith individual, A_i is the q-dimentional vector of random effects which is identically and independently distributed as $\mathbb{N}_q(\boldsymbol{\alpha}, \Sigma)$, $\boldsymbol{x}_i$ is the random effects covariate vector and E_i is the N_i-dimensional vector of errors which is identically and independently distributed as $\mathbb{N}_{N_i}(\mathbf{0}, \sigma^2\boldsymbol{I}_{N_i})$, where $\boldsymbol{I}_{N_i}$ is the N_i-dimensional identity matrix. The observations from the new individual is modeled as a nonlinear random effects model

$$Y_0 = g_0(A_0, \boldsymbol{x}_0) + E_0, \tag{4.27}$$

with

$$Y_0 = \begin{pmatrix} Y_0^P \\ Y_0^F \end{pmatrix}, \quad \boldsymbol{x}_0 = \begin{pmatrix} \boldsymbol{x}_0^P \\ x_0^F \end{pmatrix} \quad E_0 = \begin{pmatrix} E_0^P \\ E_0^F \end{pmatrix}, \quad g_0(A_0, \boldsymbol{x}_0) : \mathbb{R}^q \times \mathbb{R}^{N_0+1} \to \mathbb{R}^{N_0+1},$$

$$g_0(A_0, \boldsymbol{x}_0) = \begin{pmatrix} g_0^P(A_0, \boldsymbol{x}_0^P) \\ g(A_0, x_0^F) \end{pmatrix}, \quad g_0^P(A_0, \boldsymbol{x}_0^P) = \begin{pmatrix} g(A_0, x_{01}) \\ \vdots \\ g(A_0, x_{0N_0}) \end{pmatrix}$$

and

$$A_0 = \boldsymbol{\alpha} + \boldsymbol{\delta}_0, \quad \boldsymbol{\delta}_0 \sim \mathbb{N}_q(\mathbf{0}, \Sigma), \quad E_0 \sim \mathbb{N}_{N_0+1}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{N_0+1})$$

where indices P and F show respectively the observed (Past) and unobserved (Future) components of the new individual and $\boldsymbol{I}_{N_0+1}$ is the (N_0+1)-dimensional identity matrix. It is also supposed that $A_0, A_1, \cdots, A_I, E_0^P, E_0^F, E_1, \cdots, E_I$ are independent. The unknown observation of the new individual at x_0^F is modeled as

$$Y_0^F = g(A_0, x_0^F) + E_0^F, \tag{4.28}$$

where $A_0 \sim \mathbb{N}(\boldsymbol{\alpha}, \Sigma)$ and $E_0^F \sim \mathbb{N}(0, \sigma^2)$.

We extend the Liski and Nummi's estimation and prediction procedures for linear random effects modesl to nonlinear ones by linearization of the nonlinear function around the starting values of A_i and then apply the obtained pseudo data in the related estimation and prediction procedures. Take the first-order Taylor expansion of $g_i(A_i, \boldsymbol{x}_i), \quad i \in \{1, \cdots, I\}$ and $g_0(A_0, \boldsymbol{x}_0)$ around the starting values $\tilde{a}_i$ and $\tilde{a}_0$ (which are obtained from the Pinheiro and Bates' estimation procedure) as

$$Y_i \approx g_i(\tilde{a}_i, \boldsymbol{x}_i) + \tilde{\boldsymbol{X}}_i(A_i - \tilde{a}_i) + E_i, \quad i \in \{1, \cdots, I\}$$

and

$$Y_0 \approx g_0(\tilde{a}_0, \boldsymbol{x}_0) + \tilde{\boldsymbol{X}}_0(A_0 - \tilde{a}_0) + E_0,$$

where

$$\tilde{\boldsymbol{X}}_i := \frac{\partial g_i(a_i, \boldsymbol{x}_i)}{\partial a_i}|_{a_i = \tilde{a}_i}, \in \mathbb{R}^{N_i \times q}, \quad \tilde{\boldsymbol{X}}_0 := \frac{\partial g_0(a_0, \boldsymbol{x}_0)}{\partial a_0}|_{a_0 = \tilde{a}_0} \in \mathbb{R}^{(N_0+1) \times q}.$$

Define the pseudo data as

$$Z_i := Y_i - g_i(\tilde{a}_i, \boldsymbol{x}_i) + \tilde{\boldsymbol{X}}_i \tilde{a}_i \approx \tilde{\boldsymbol{X}}_i A_i + E_i, \quad i \in \{1, \cdots, I\} \tag{4.29}$$

and

$$Z_0 := Y_0 - g_0(\tilde{a}_0, \boldsymbol{x}_0) + \tilde{\boldsymbol{X}}_0 \tilde{a}_0 \approx \tilde{\boldsymbol{X}}_0 A_0 + E_0, \tag{4.30}$$

where

$$Z_0 = \begin{pmatrix} Z_0^P \\ Z_0^F \end{pmatrix}, \quad \tilde{\boldsymbol{X}}_0 = \begin{pmatrix} \tilde{\boldsymbol{X}}_0^P \\ \tilde{\boldsymbol{x}}_0^F \end{pmatrix}, \quad g_0(\tilde{a}_0, \boldsymbol{x}_0) = \begin{pmatrix} g_0^P(\tilde{a}_0, \boldsymbol{x}_0^P) \\ g(\tilde{a}_0, x_0^F) \end{pmatrix}, \quad E_0 = \begin{pmatrix} E_0^P \\ E_0^F \end{pmatrix},$$

and

$$\tilde{\boldsymbol{X}}_0^P := \frac{\partial g_0^P(a_0, \boldsymbol{x}_0^P)}{\partial a_0}|_{a_0=\tilde{a}_0}, \quad \tilde{\boldsymbol{x}}_0^F := \frac{\partial g(a_0, x_0^F)}{\partial a_0}|_{a_0=\tilde{a}_0}.$$

Based on the constructed linear random effects models (4.29) and (4.30), the estimation and prediction methods of Liski and Nummi (1996) can be applied for the observed pseudo data $\boldsymbol{Z}_{obs} := (Z_1^\intercal, \cdots, Z_I^\intercal, Z_0^{P\intercal})^\intercal$. If the variance components are known, then the prediction of Z_0^F is given by

$$\tilde{Z}_0^F := \tilde{\boldsymbol{x}}_0^F \tilde{\boldsymbol{\alpha}} + V_0^{FP}(V_0^P)^{-1}(Z_0^P - \tilde{\boldsymbol{X}}_0^P \tilde{\boldsymbol{\alpha}}), \tag{4.31}$$

where

$$V_0^{FP} := \mathbb{C}\mathrm{ov}(Z_0^F, Z_0^P) = \tilde{\boldsymbol{x}}_0^F \Sigma (\tilde{\boldsymbol{X}}_0^P)^\intercal,$$

and

$$\tilde{\boldsymbol{\alpha}} := [\sum_{i=1}^{I} \tilde{\boldsymbol{X}}_i^\intercal V_i^{-1} \tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal} V_0^{P^{-1}} \tilde{\boldsymbol{X}}_0^P]^{-1} (\sum_{i=1}^{I} \tilde{\boldsymbol{X}}_i^\intercal V_i^{-1} Z_i + \tilde{\boldsymbol{X}}_0^{P\intercal} V_0^{P^{-1}} Z_0^P),$$

with

$$V_i := \mathbb{C}\mathrm{ov}(Z_i) = \tilde{\boldsymbol{X}}_i \Sigma \tilde{\boldsymbol{X}}_i^\intercal + \sigma^2 \boldsymbol{I}_{N_i}, \quad i \in \{1, \cdots, I\}$$

and

$$V_0^P := \mathbb{C}\mathrm{ov}(Z_0^P) = \tilde{\boldsymbol{X}}_0^P \Sigma \tilde{\boldsymbol{X}}_0^{P\intercal} + \sigma^2 \boldsymbol{I}_{N_0}.$$

By use of the relationship between the pseudo data and real data (4.30) we have

$$\tilde{Y}_0^F = \tilde{Z}_0^F + g(\tilde{a}_0, x_0^F) - \tilde{\boldsymbol{x}}_0^F \tilde{a}_0, \tag{4.32}$$

$$Y_0^F = Z_0^F + g(\tilde{a}_0, x_0^F) - \tilde{\boldsymbol{x}}_0^F \tilde{a}_0. \tag{4.33}$$

According to (4.32) and (4.33), it is readily concluded that

$$\tilde{Z}_0^F - Z_0^F = \tilde{Y}_0^F - Y_0^F. \tag{4.34}$$

By use of (4.34) and Lemma 4.3, the variance of prediction error $\tilde{Y}_0^F - Y_0^F$ is given by

$$\mathbb{V}\mathrm{ar}[\tilde{Y}_0^F - Y_0^F] = \mathbb{V}\mathrm{ar}[\tilde{Z}_0^F - Z_0^F] = \Omega_0^F + M_0[\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})]M_0^\intercal =: \Omega_0,$$

where

$$\Omega_0^F = V_0^F - V_0^{FP}(V_0^P)^{-1}(V_0^{FP})^\intercal, \quad M_0 = \tilde{\boldsymbol{x}}_0^F - V_0^{FP}(V_0^P)^{-1}\tilde{\boldsymbol{X}}_0^P,$$

with

$$V_0^F := \mathbb{V}\mathrm{ar}(Z_0^F) = \tilde{\boldsymbol{x}}_0^F \Sigma \tilde{\boldsymbol{x}}_0^{F\intercal} + \sigma^2, \quad V_0^P := \mathbb{C}\mathrm{ov}(Z_0^P) = \tilde{\boldsymbol{X}}_0^P \Sigma \tilde{\boldsymbol{X}}_0^{P\intercal} + \sigma^2 \boldsymbol{I}_{N_0},$$

$$V_0^{FP} := \mathbb{C}\mathrm{ov}(Z_0^F, Z_0^P) = \tilde{\boldsymbol{x}}_0^F \Sigma \tilde{\boldsymbol{X}}_0^{P\intercal}, \ \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}}) = [\sum_{i=1}^{I} \tilde{X}_i^\intercal V_i^{-1} \tilde{X}_i + \tilde{X}_0^{P\intercal} V_0^{P^{-1}} \tilde{X}_0^P]^{-1}.$$

In final, the $100(1-\alpha)\%$ prediction interval for Y_0^F is given by:

$$\tilde{y}_0^F \pm q_{1-\alpha/2} \sqrt{\Omega_0},$$

where $\tilde{y}_0^F$ is the realization of $\tilde{Y}_0^F$. If the variance components are unknown, then the achieved ML estimation $\hat{\boldsymbol{\theta}} := (\hat{\sigma}^2, \mathrm{vech}^\intercal(\hat{\Sigma}))^\intercal$, which is obtained from the Liski and Nummi's estimation by use of the pseudo data, is replaced in the prediction error variance expression. If $\hat{\boldsymbol{\alpha}}$ is the obtained ML estimate from the Liski and Nummi's estimation by use of the pseudo data, then the prediction of unknown observation from the partially observed individual is given by

$$\hat{y}_0^F = \hat{z}_0^F + g(\tilde{a}_0, x_0^F) - \tilde{\boldsymbol{x}}_0^F \tilde{a}_0,$$

where

$$\hat{z}_0^F := \tilde{\boldsymbol{x}}_0^F \hat{\boldsymbol{\alpha}} + V_0^{FP}(\hat{\boldsymbol{\theta}}) V_0^{P^{-1}}(\hat{\boldsymbol{\theta}})(z_0^P - \tilde{\boldsymbol{X}}_0^P \hat{\boldsymbol{\alpha}}),$$

with

$$V_0^{FP}(\hat{\boldsymbol{\theta}}) := \tilde{\boldsymbol{x}}_0^F \hat{\Sigma} \tilde{\boldsymbol{X}}_0^{P\intercal}, \quad V_0^P(\hat{\boldsymbol{\theta}}) := \tilde{\boldsymbol{X}}_0^P \hat{\Sigma} \tilde{\boldsymbol{X}}_0^{P\intercal} + \hat{\sigma}^2 \boldsymbol{I}_{N_0},$$

and z_0^P is the realization of Z_0^P.

4.2.2 Combination of method of Pinheiro/Bates and the modified method of Swamy in the unbalanced case (PBmSwamy)

In this method, we extend the modified method of Swamy's prediction for linear models to nonlinear ones, by use of the same linearization procedure which has been proposed in the last section such that the starting values $\tilde{a}_i$ for $i = 1, \cdots, I$ and $\tilde{a}_0$ are obtained from the Pinheiro and Bates' estimation procedure. By use of the relationship between the pseudo data and real data (4.30) we have

$$\hat{Y}_0^F = \hat{Z}_0^F + g(\tilde{a}_0, x_0^F) - \tilde{\boldsymbol{x}}_0^F \tilde{a}_0, \tag{4.35}$$

$$Y_0^F = Z_0^F + g(\tilde{a}_0, x_0^F) - \tilde{\boldsymbol{x}}_0^F \tilde{a}_0. \tag{4.36}$$

According to (4.35) and (4.36), it is readily concluded that

$$\hat{Z}_0^F - Z_0^F = \hat{Y}_0^F - Y_0^F. \tag{4.37}$$

In the modified method of Swamy's prediction for linear random effects models (4.29) and (4.30) in the unbalnced case, we predict Z_0^F as follows:

$$\hat{Z}_0^F = (1-w)\hat{Z}_0^{F(old)} + w\hat{Z}_0^{F(new)},$$

where

$$\hat{Z}_0^{F(old)} = \tilde{\boldsymbol{x}}_0^F \hat{\boldsymbol{\alpha}}, \quad \hat{\boldsymbol{\alpha}} = \frac{1}{I}\sum_{i=1}^{I} \hat{A}_{i,OLS} = \frac{1}{I}\sum_{i=1}^{I} (\tilde{\boldsymbol{X}}_i^\intercal \tilde{\boldsymbol{X}}_i)^{-1} \tilde{\boldsymbol{X}}_i^\intercal Z_i,$$

and

$$\hat{Z}_0^{F(new)} = \tilde{\boldsymbol{x}}_0^F \hat{A}_{0,OLS} = \tilde{\boldsymbol{x}}_0^F (\tilde{\boldsymbol{X}}_0^{P\intercal} \tilde{\boldsymbol{X}}_0^P)^{-1} \tilde{\boldsymbol{X}}_0^{P\intercal} Z_0^P$$

with

$$\tilde{\boldsymbol{X}}_i := \frac{\partial g_i(a_i, \boldsymbol{x}_i)}{\partial a_i}|_{a_i=\tilde{a}_i}, \quad \tilde{\boldsymbol{X}}_0^P := \frac{\partial g_0^P(a_0, \boldsymbol{x}_0^P)}{\partial a_0}|_{a_0=\tilde{a}_0},$$

$$\tilde{\boldsymbol{x}}_0^F = \frac{\partial g(a_0, x_0^F)}{\partial a_0}|_{a_0=\tilde{a}_0} \in \mathbb{R}^q, \quad w = \frac{k}{k+h},$$

$$k := \frac{1}{I}\tilde{\boldsymbol{x}}_0^F (\Sigma + \frac{\sigma^2}{I}\sum_{i=1}^{I} (\tilde{\boldsymbol{X}}_i^\intercal \tilde{\boldsymbol{X}}_i)^{-1}) \tilde{\boldsymbol{x}}_0^{F\intercal} + \tilde{\boldsymbol{x}}_0^F \Sigma \tilde{\boldsymbol{x}}_0^{F\intercal}, \quad h := \sigma^2 \tilde{\boldsymbol{x}}_0^F (\tilde{\boldsymbol{X}}_0^{P\intercal} \tilde{\boldsymbol{X}}_0^P)^{-1} \tilde{\boldsymbol{x}}_0^{F\intercal}.$$

In Lemma 4.2, the distribution of prediction error for linear models in the balanced case is presented. This lemma can be easily extended to the unbalanced case such that the prediction error has normal distribution with mean 0 and variance $\nu := (1-w)^2 k + w^2 h + \sigma^2$. Because, σ^2 and Σ are unknown we need to estimate them as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{I} (Z_i - \tilde{\boldsymbol{X}}_i \hat{A}_{i,OLS})^\intercal (Z_i - \tilde{\boldsymbol{X}}_i \hat{A}_{i,OLS})}{\sum_{i=1}^{I} (N_i - q)},$$

$$\hat{\Sigma} = \frac{1}{I-1}\sum_{i=1}^{I} (\hat{A}_{i,OLS} - \hat{\boldsymbol{\alpha}})(\hat{A}_{i,OLS} - \hat{\boldsymbol{\alpha}})^\intercal - \frac{\hat{\sigma}^2}{I}\sum_{i=1}^{I} (\tilde{\boldsymbol{X}}_i^\intercal \tilde{\boldsymbol{X}}_i)^{-1},$$

with $\hat{A}_{i,OLS} = (\tilde{\boldsymbol{X}}_i^\intercal \tilde{\boldsymbol{X}}_i)^{-1} \tilde{\boldsymbol{X}}_i^\intercal Z_i$ (Carter and Yang, 1986).
Hence, from (4.37) and Lemma 4.2 we have $\hat{Y}_0^F - Y_0^F \sim \mathbb{N}(0, \nu)$. ν can be estimated by $\hat{\nu}$ as follows:

$$\hat{\nu} = (1-\hat{w})^2 \hat{k} + \hat{w}^2 \hat{h} + \hat{\sigma}^2,$$

where $\hat{w} = \frac{\hat{k}}{\hat{k}+\hat{h}}$ and

$$\hat{k} = \frac{1}{I}\tilde{\boldsymbol{x}}_0^F (\hat{\Sigma} + \frac{\hat{\sigma}^2}{I}\sum_{i=1}^{I} (\tilde{\boldsymbol{X}}_i^\intercal \tilde{\boldsymbol{X}}_i)^{-1}) \tilde{\boldsymbol{x}}_0^{F\intercal} + \tilde{\boldsymbol{x}}_0^F \hat{\Sigma} \tilde{\boldsymbol{x}}_0^{F\intercal},$$

$$\hat{h} = \hat{\sigma}^2 \tilde{\boldsymbol{x}}_0^F (\tilde{\boldsymbol{X}}_0^{P\intercal} \tilde{\boldsymbol{X}}_0^P)^{-1} \tilde{\boldsymbol{x}}_0^{F\intercal}.$$

In final, the $100(1-\alpha)\%$ prediction interval for Y_0^F is given by:

$$\hat{y}_0^F \pm q_{1-\alpha/2}\sqrt{\hat{\nu}},$$

where $\hat{y}_0^F$ is the realization of $\hat{Y}_0^F$ and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of standard normal distribution.

4.2.3 The modified method of Pinheiro and Bates (mPB)

Consider the defined nonlinear random effects model (4.26) for the completely observed individuals and the nonlinear random effects model (4.27) for the partially observed individual (new individual) with the proposed decomposition. Let the unknown observation from the new individual is modeled as (4.28). Estimation of the expected response, when the random effect ($\boldsymbol{\delta}_0$) is equal to its mean value $\mathbf{0}$, is a simple plug-in predictor for Y_0^F (Pinheiro and Bates, 2000). The proposed simple plug-in prediction of Y_0^F when $\boldsymbol{\delta}_0 = \mathbf{0}$ is given by

$$\hat{Y}_0^F := g(\hat{\boldsymbol{\alpha}}, x_0^F), \tag{4.38}$$

where $\hat{\boldsymbol{\alpha}}$ is the maximum likelihood estimation of $\boldsymbol{\alpha}$ based on the information of $\boldsymbol{Y}_{obs} := (Y_1^\intercal, \cdots, Y_I^\intercal, Y_0^{P\intercal})^\intercal$. Prediction (4.38) is a modified version of Pinheiro and Bates' prediction procedure because for the estimation of $\boldsymbol{\alpha}$ the information of the new individual is also considered, whereas in the original one only the information of the old individuals is considered. The ML estimates of $\boldsymbol{\alpha}$, Σ and σ^2 are obtained from the estimation procedure proposed by Pinheiro and Bates (2000) on the basis of the information of $\boldsymbol{Y}_{obs}$ (see section 3.2.1 for more details).
In the following theorem, we propose an approximation method for finding distribution of the prediction error for the modified Pinheiro and Bates' prediction and finally an approximate prediction interval.

Theorem 4.2. *Approximate prediction interval for Y_0^F based on the modified Pinheiro and Bates' prediction:*
Consider the nonlinear random effects model (4.26) and the defined nonlinear random effects model for the new individual (4.27). If we consider the simple plug-in predictor of Y_0^F as $\hat{Y}_0^F := g(\hat{\boldsymbol{\alpha}}, x_0^F)$ then the prediction error $(Y_0^F - \hat{Y}_0^F)$ has approximate normal distribution with the mean 0 and estimated variance

$$\hat{\nu} := \tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^F [\sum_{i=1}^{I} \tilde{\boldsymbol{X}}_i^\intercal \widehat{V}_i^{-1} \tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal} \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}_0^P]^{-1} [\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal} - 2\tilde{\boldsymbol{X}}_0^{P\intercal} \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}_0^P \hat{\Sigma} \tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal}]$$

$$+\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F}\hat{\Sigma}\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal}+\hat{\sigma}^2,$$

where

$$\tilde{\boldsymbol{X}}_i := \frac{\partial g_i(\boldsymbol{a},\boldsymbol{x}_i)}{\partial \boldsymbol{a}^\intercal}|_{\boldsymbol{a}=\hat{\boldsymbol{\alpha}}+\hat{\boldsymbol{\delta}}_i}, \quad \tilde{\boldsymbol{X}}_0^P := \frac{\partial g_0^P(\boldsymbol{a},\boldsymbol{x}_0^P)}{\partial \boldsymbol{a}^\intercal}|_{\boldsymbol{a}=\hat{\boldsymbol{\alpha}}+\hat{\boldsymbol{\delta}}_0},$$

$$\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F} := \frac{\partial g(\boldsymbol{a},x_0^F)}{\partial \boldsymbol{a}^\intercal}|_{\boldsymbol{a}=\hat{\boldsymbol{\alpha}}},$$

with

$$\widehat{V_i} := \tilde{\boldsymbol{X}}_i\hat{\Sigma}\tilde{\boldsymbol{X}}_i^\intercal+\hat{\sigma}^2\boldsymbol{I}_{N_i}, \quad \widehat{V_0}^P := \tilde{\boldsymbol{X}}_0^P\hat{\Sigma}\tilde{\boldsymbol{X}}_0^{P\intercal}+\hat{\sigma}^2\boldsymbol{I}_{N_0},$$

and $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\delta}}_i$, $\hat{\Sigma}=\hat{\sigma}^2(\Delta^\intercal(\hat{\boldsymbol{\theta}})\Delta(\hat{\boldsymbol{\theta}}))^{-1}$ *and* $\hat{\sigma}^2$ *are the obtained estimated values from the Pinheiro and Bates' estimation method based on the information of* $\boldsymbol{Y}_{obs}$*. Moreover, the* $(1-\alpha)$*-prediction interval for* Y_0^F *from a partially observed individual is given by*

$$[\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}], \tag{4.39}$$

where $\hat{y}_0^F$ *is the realization of* $\hat{Y}_0^F$ *and* $q_{1-\alpha/2}$ *is the* $(1-\alpha/2)$*-quantile of standard normal distribution.*

Proof. If we consider the first order Taylor expansion of $g(A_0,x_0^F)$ and $g(\hat{\boldsymbol{\alpha}},x_0^F)$ around $\boldsymbol{\alpha}$, then the prediction error can be approximated as

$$\begin{aligned}
&Y_0^F-\hat{Y}_0^F = g(A_0,x_0^F)+E_0^F-g(\hat{\boldsymbol{\alpha}},x_0^F)\\
&\approx g(\boldsymbol{\alpha},x_0^F)+\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F\boldsymbol{\delta}_0-g(\boldsymbol{\alpha},x_0^F)-\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})+E_0^F\\
&=\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}})+\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F\boldsymbol{\delta}_0+E_0^F,
\end{aligned} \tag{4.40}$$

where $\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F = \frac{\partial g(\boldsymbol{a},x_0^F)}{\partial \boldsymbol{a}^\intercal}|_{\boldsymbol{a}=\boldsymbol{\alpha}}$. Pinheiro and Bates (2000) prove that $\hat{\boldsymbol{\alpha}}$ has approximately normal distribution with mean $\boldsymbol{\alpha}$ and the following estimated variance covariance matrix:

$$\mathbb{E}(\hat{\boldsymbol{\alpha}})\approx\boldsymbol{\alpha}, \quad \widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}})\approx[\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^\intercal\widehat{V}_i^{-1}\tilde{\boldsymbol{X}}_i+\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^P]^{-1}. \tag{4.41}$$

From (4.41), because $\hat{\boldsymbol{\alpha}}$ has normal distribution hence $Y_0^F-\hat{Y}_0^F$, which is approximately a linear combination of the three normal variables $(\hat{\boldsymbol{\alpha}},\boldsymbol{\delta}_0,E_0^F)$, has normal distribution. From (4.41), we have $\mathbb{E}(\hat{\boldsymbol{\alpha}})\approx\boldsymbol{\alpha}$, then the expectation of prediction error is given by

$$\mathbb{E}(Y_0^F-\hat{Y}_0^F)\approx\mathbb{E}(\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}})+\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F\boldsymbol{\delta}_0+E_0^F)$$

$$
\begin{aligned}
&= \tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}(\boldsymbol{\alpha} - \mathbb{E}(\hat{\boldsymbol{\alpha}})) + \tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}\mathbb{E}(\boldsymbol{\delta}_0) + \mathbb{E}(E^F_0) \\
&\approx \tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}(\boldsymbol{\alpha} - \boldsymbol{\alpha}) + \tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}.\mathbf{0} + 0 \\
&= 0.
\end{aligned}
$$

By use of the independency between $\boldsymbol{\delta}_0$ and E^F_0 and between E^F_0 and $\hat{\boldsymbol{\alpha}}$, the estimated variance of prediction error is given by

$$
\begin{aligned}
\widehat{\mathbb{V}\mathrm{ar}}(Y^F_0 - \hat{Y}^F_0) &\approx \widehat{\mathbb{V}\mathrm{ar}}(-\tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}} + \tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}\boldsymbol{\delta}_0 + E^F_0) \\
&= \widehat{\mathbb{V}\mathrm{ar}}(\tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}) + \widehat{\mathbb{V}\mathrm{ar}}(\tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}\boldsymbol{\delta}_0 + E^F_0) - 2\widehat{\mathbb{C}\mathrm{ov}}(\tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}, \tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}\boldsymbol{\delta}_0 + E^F_0) \\
&= \tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}})\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}} + \widehat{\mathbb{V}\mathrm{ar}}(E^F_0) + \tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\hat{\Sigma}\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}} \\
&\quad + 2\tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\widehat{\mathbb{C}\mathrm{ov}}(\boldsymbol{\delta}_0, E^F_0) - 2[\tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\delta}_0)\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}} + \tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}}, E^F_0)] \\
&= \tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}})\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}} + \hat{\sigma}^2 + \tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\hat{\Sigma}\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}} - 2\tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\delta}_0)\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}}. \qquad (4.42)
\end{aligned}
$$

As it was discussed in Section 3.2.1, the maximum likelihood estimation of $\boldsymbol{\alpha}$ by use of the information of $\boldsymbol{Y}_{obs}$ is given by

$$
\hat{\boldsymbol{\alpha}} = [\sum_{i=1}^{I} \tilde{\boldsymbol{X}}^{\intercal}_i \widehat{V_i}^{-1} \tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}^{P\intercal}_0 \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}^P_0]^{-1} (\sum_{i=1}^{I} \tilde{\boldsymbol{X}}^{\intercal}_i \widehat{V_i}^{-1} Z_i + \tilde{\boldsymbol{X}}^{P\intercal}_0 \widehat{V_0}^{P^{-1}} Z^P_0)
$$

with

$$
\begin{aligned}
Z_i &:= Y_i - g_i(\hat{a}_i, \boldsymbol{x}_i) + \tilde{\boldsymbol{X}}_i\hat{a}_i \approx \tilde{\boldsymbol{X}}_i A_i + E_i, \\
Z^P_0 &:= Y^P_0 - g^P_0(\hat{a}_0, \boldsymbol{x}^P_0) + \tilde{\boldsymbol{X}}^P_0\hat{a}_0 \approx \tilde{\boldsymbol{X}}^P_0 A_0 + E^P_0,
\end{aligned}
$$

where $\hat{a}_i = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\delta}}_i$ is the estimated value of A_i for $i = 0, 1, \cdots, I$.
Therefore by using the independency between E^P_0, $\boldsymbol{\delta}_0$ and E^F_0, we have

$$
\tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\delta}_0)\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}} = \tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}[\sum_{i=1}^{I} \tilde{\boldsymbol{X}}^{\intercal}_i \widehat{V_i}^{-1} \tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}^{P\intercal}_0 \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}^P_0]^{-1} \tilde{\boldsymbol{X}}^{P\intercal}_0 \widehat{V_0}^{P^{-1}} \widehat{\mathbb{C}\mathrm{ov}}(Z^P_0, \boldsymbol{\delta}_0)\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}}
$$

$$
= \tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}[\sum_{i=1}^{I} \tilde{\boldsymbol{X}}^{\intercal}_i \widehat{V_i}^{-1} \tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}^{P\intercal}_0 \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}^P_0]^{-1} \tilde{\boldsymbol{X}}^{P\intercal}_0 \widehat{V_0}^{P^{-1}} \widehat{\mathbb{C}\mathrm{ov}}(\tilde{\boldsymbol{X}}^P_0 A_0 + E^P_0, \boldsymbol{\delta}_0)\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}}
$$

$$
= \tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}[\sum_{i=1}^{I} \tilde{\boldsymbol{X}}^{\intercal}_i \widehat{V_i}^{-1} \tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}^{P\intercal}_0 \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}^P_0]^{-1} \tilde{\boldsymbol{X}}^{P\intercal}_0 \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}^P_0 \hat{\Sigma}\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}},
$$

From (4.41) and (4.42), the estimated variance of prediction error is given by

$$
\widehat{\mathbb{V}\mathrm{ar}}(Y^F_0 - \hat{Y}^F_0) \approx \tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}})\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}} + \tilde{\boldsymbol{x}}^F_{0,\hat{\boldsymbol{\alpha}}}\hat{\Sigma}\tilde{\boldsymbol{x}}^{F\intercal}_{0,\hat{\boldsymbol{\alpha}}} + \hat{\sigma}^2
$$

$$
\begin{aligned}
&-2\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F}[\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^{\intercal}\widehat{V}_i^{-1}\tilde{\boldsymbol{X}}_i+\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V}_0^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}]^{-1}\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V}_0^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}\hat{\Sigma}\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal}\\
&\approx\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F}[\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^{\intercal}\widehat{V}_i^{-1}\tilde{\boldsymbol{X}}_i+\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V}_0^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}]^{-1}\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal}+\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F}\hat{\Sigma}\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal}+\hat{\sigma}^2\\
&\quad-2\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F}[\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^{\intercal}\widehat{V}_i^{-1}\tilde{\boldsymbol{X}}_i+\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V}_0^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}]^{-1}\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V}_0^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}\hat{\Sigma}\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal}\\
&=\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F}[\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^{\intercal}\widehat{V}_i^{-1}\tilde{\boldsymbol{X}}_i+\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V}_0^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}]^{-1}[\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal}-2\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V}_0^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}\hat{\Sigma}\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal}]\\
&+\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F}\hat{\Sigma}\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal}+\hat{\sigma}^2\\
&=:\hat{\nu}
\end{aligned}
$$

In conclusion,

$$Y_0^F-\hat{Y}_0^F\sim\mathbb{N}(0,\hat{\nu}),$$

and then the $(1-\alpha)$-prediction interval for Y_0^F from a partially observed individual is given by

$$[\hat{y}_0^F-q_{1-\alpha/2}\sqrt{\hat{\nu}},\hat{y}_0^F+q_{1-\alpha/2}\sqrt{\hat{\nu}}],$$

□

4.2.4 Prediction approach of Hall and Clutter

Consider the nonlinear random effects models (4.26) and (4.27). Let $\boldsymbol{\beta}:=(\boldsymbol{\alpha},\Sigma,\sigma^2)$ and the ML estimate $\hat{\boldsymbol{\beta}}:=(\hat{\boldsymbol{\alpha}},\hat{\Sigma},\hat{\sigma}^2)$ be obtained based on the information of $\boldsymbol{Y}_{obs}$ from the Pinheiro and Bates' estimation method. Hall and Clutter (2004) propose a predictor based on the linear approximation of the nonlinear model. In the following, Hall and Clutter's prediction procedure is presented.
By taking the first order Taylor expansion of $g_0(A_0,\boldsymbol{x}_0)$ around the estimated value $\hat{a}_0=\hat{\boldsymbol{\alpha}}+\hat{\boldsymbol{\delta}}_0$, we have

$$Y_0\approx g_0(\hat{a}_0,\boldsymbol{x}_0)+\tilde{\boldsymbol{X}}_0(A_0-\hat{a}_0)+E_0,\tag{4.43}$$

where

$$\tilde{\boldsymbol{X}}_0:=\frac{\partial g_0(a_0,\boldsymbol{x}_0)}{\partial a_0}|_{a_0=\hat{a}_0}\in\mathbb{R}^{(N_0+1)\times q}.$$

Define the pseudo data as

$$Z_0:=Y_0-g_0(\hat{a}_0,x_0)+\tilde{\boldsymbol{X}}_0\hat{a}_0\approx\tilde{\boldsymbol{X}}_0A_0+E_0,\tag{4.44}$$

where A_0 can be represented as $A_0 = \boldsymbol{\alpha} + \boldsymbol{\delta}_0$ and

$$Z_0 = \begin{pmatrix} Z_0^P \\ Z_0^F \end{pmatrix}, \quad \tilde{\boldsymbol{X}}_0 = \begin{pmatrix} \tilde{\boldsymbol{X}}_0^P \\ \tilde{\boldsymbol{x}}_0^F \end{pmatrix}, \quad g_0(\hat{a}_0, \boldsymbol{x}_0) = \begin{pmatrix} g_0^P(\hat{a}_0, \boldsymbol{x}_0^P) \\ g(\hat{a}_0, x_0^F) \end{pmatrix}, \quad E_0 = \begin{pmatrix} E_0^P \\ E_0^F \end{pmatrix}.$$

Based on the linear random effects model (4.44), the emperical BLUP of Z_0^F based on Z_0^P is given by

$$\hat{Z}_0^F = \mathbb{E}(Z_0^F | Z_0^P, \hat{\boldsymbol{\beta}}) = \tilde{\boldsymbol{x}}_0^F \hat{\boldsymbol{\alpha}} + \widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}} (Z_0^P - \tilde{\boldsymbol{X}}_0^P \hat{\boldsymbol{\alpha}}), \tag{4.45}$$

where

$$\widehat{V_0}^{FP} := \widehat{\mathbb{C}\mathrm{ov}}(Z_0^F, Z_0^P) = \tilde{\boldsymbol{x}}_0^F \hat{\Sigma} \tilde{\boldsymbol{X}}_0^{P\intercal}, \quad \widehat{V_0}^{P} := \widehat{\mathbb{C}\mathrm{ov}}(Z_0^P) = \tilde{\boldsymbol{X}}_0^P \hat{\Sigma} \tilde{\boldsymbol{X}}_0^{P\intercal} + \hat{\sigma}^2 \boldsymbol{I}_{N_0},$$

$$\hat{\boldsymbol{\alpha}} = [\sum_{i=1}^{I} \tilde{\boldsymbol{X}}_i^\intercal \widehat{V_i}^{-1} \tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal} \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}_0^P]^{-1} (\sum_{i=1}^{I} \tilde{\boldsymbol{X}}_i^\intercal \widehat{V_i}^{-1} Z_i + \tilde{\boldsymbol{X}}_0^{P\intercal} \widehat{V_0}^{P^{-1}} Z_0^P), \tag{4.46}$$

with

$$Z_i := Y_i - g_i(\hat{a}_i, \boldsymbol{x}_i) + \tilde{\boldsymbol{X}}_i \hat{a}_i \approx \tilde{\boldsymbol{X}}_i A_i + E_i, \tag{4.47}$$

$$Z_0^P := Y_0^P - g_0^P(\hat{a}_0, \boldsymbol{x}_0^P) + \tilde{\boldsymbol{X}}_0^P \hat{a}_0 \approx \tilde{\boldsymbol{X}}_0^P A_0 + E_0^P,$$

$$\tilde{\boldsymbol{X}}_i := \frac{\partial g_i(a_i, \boldsymbol{x}_i)}{\partial a_i}|_{a_i = \hat{a}_i}, \quad \tilde{\boldsymbol{X}}_0^P := \frac{\partial g_0^P(a_0, \boldsymbol{x}_0^P)}{\partial a_0}|_{a_0 = \hat{a}_0}, \quad \tilde{\boldsymbol{x}}_0^F := \frac{\partial g(a_0, x_0^F)}{\partial a_0}|_{a_0 = \hat{a}_0},$$

$$\widehat{V_i} = \widehat{\mathbb{C}\mathrm{ov}}(Z_i) = \tilde{\boldsymbol{X}}_i \hat{\Sigma} \tilde{\boldsymbol{X}}_i^\intercal + \hat{\sigma}^2 \boldsymbol{I}_{N_i},$$

and $\hat{a}_i$, $i \in \{0, 1, \cdots, I\}$ is the estimated value of A_i from the Pinheiro and Bates' estimation method.
From (4.44) and (4.45) we have

$$\hat{Y}_0^F - g(\hat{a}_0, x_0^F) + \tilde{\boldsymbol{x}}_0^F \hat{a}_0 = \tilde{\boldsymbol{x}}_0^F \hat{\boldsymbol{\alpha}} + \widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}} (Z_0^P - \tilde{\boldsymbol{X}}_0^P \hat{\boldsymbol{\alpha}})$$

$$\iff \hat{Y}_0^F = g(\hat{a}_0, x_0^F) - \tilde{\boldsymbol{x}}_0^F \hat{a}_0 + \tilde{\boldsymbol{x}}_0^F \hat{\boldsymbol{\alpha}} + \widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}} (Z_0^P - \tilde{\boldsymbol{X}}_0^P \hat{\boldsymbol{\alpha}}).$$

By use of $\hat{a}_0 = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\delta}}_0$, it is concluded that

$$\hat{Y}_0^F = g(\hat{a}_0, x_0^F) - \tilde{\boldsymbol{x}}_0^F \hat{\boldsymbol{\delta}}_0 + \widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}} (Z_0^P - \tilde{\boldsymbol{X}}_0^P \hat{\boldsymbol{\alpha}}). \tag{4.48}$$

From (4.43) we have

$$Y_0^F \approx g(\hat{a}_0, x_0^F) + \tilde{\boldsymbol{x}}_0^F (A_0 - \hat{a}_0) + E_0^F. \tag{4.49}$$

Finally based on (4.48), (4.49) and (4.44) we have

$$
\begin{aligned}
&Y_0^F - \hat{Y}_0^F \approx g(\hat{a}_0, x_0^F) + \tilde{\boldsymbol{x}}_0^F (A_0 - \hat{a}_0) + E_0^F - g(\hat{a}_0, x_0^F) + \tilde{\boldsymbol{x}}_0^F \hat{\boldsymbol{\delta}}_0 - \widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}} (Z_0^P - \tilde{\boldsymbol{X}}_0^P \hat{\boldsymbol{\alpha}}) \\
&\approx \tilde{\boldsymbol{x}}_0^F (A_0 - \hat{a}_0) + E_0^F + \tilde{\boldsymbol{x}}_0^F \hat{\boldsymbol{\delta}}_0 - \widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}} (\tilde{\boldsymbol{X}}_0^P (\boldsymbol{\alpha} + \boldsymbol{\delta}_0) + E_0^P - \tilde{\boldsymbol{X}}_0^P \hat{\boldsymbol{\alpha}}) \\
&= \tilde{\boldsymbol{x}}_0^F (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{x}}_0^F \boldsymbol{\delta}_0 + E_0^F - \widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}} (\tilde{\boldsymbol{X}}_0^P (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{X}}_0^P \boldsymbol{\delta}_0 + E_0^P) \\
&= (-\widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}}, 1)[\begin{pmatrix} \tilde{\boldsymbol{X}}_0^P \\ \tilde{\boldsymbol{x}}_0^F \end{pmatrix} (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \begin{pmatrix} \tilde{\boldsymbol{X}}_0^P \\ \tilde{\boldsymbol{x}}_0^F \end{pmatrix} \boldsymbol{\delta}_0 + \begin{pmatrix} E_0^P \\ E_0^F \end{pmatrix}] \\
&= (-\widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}}, 1)[\tilde{\boldsymbol{X}}_0 (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{X}}_0 \boldsymbol{\delta}_0 + E_0].
\end{aligned}
\tag{4.50}
$$

Theorem 4.3. *Approximate prediction interval for Y_0^F based on the Hall and Clutter's prediction [Hall and Clutter (2004)]:*
Consider the nonlinear random effects models (4.26) and (4.27). Let $\boldsymbol{\beta} := (\boldsymbol{\alpha}, \Sigma, \sigma^2)$ and the ML estimate $\hat{\boldsymbol{\beta}} := (\hat{\boldsymbol{\alpha}}, \hat{\Sigma}, \hat{\sigma}^2)$ be obtained based on the information of $\boldsymbol{Y}_{obs}$ from the Pinheiro and Bates' estimation method. The prediction error $(Y_0^F - \hat{Y}_0^F)$ has approximately normal distribution with the mean 0 and estimated variance $\hat{\nu}$,

$$
\hat{\nu} := (-\widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}}, 1) \widehat{\mathbb{C}ov}[\tilde{\boldsymbol{X}}_0 (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{X}}_0 \boldsymbol{\delta}_0 + E_0] \begin{pmatrix} (-\widehat{V_0}^{FP} \widehat{V_0}^{P^{-1}})^{\intercal} \\ 1 \end{pmatrix},
$$

where $\hat{\boldsymbol{\alpha}}$ is given by (4.46) and

$$
\begin{aligned}
&\widehat{\mathbb{C}ov}[\tilde{\boldsymbol{X}}_0 (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{X}}_0 \boldsymbol{\delta}_0 + E_0] \approx \tilde{\boldsymbol{X}}_0 \widehat{\mathbb{C}ov}(\hat{\boldsymbol{\alpha}}) \tilde{\boldsymbol{X}}_0^{\intercal} + \tilde{\boldsymbol{X}}_0 \hat{\Sigma} \tilde{\boldsymbol{X}}_0^{\intercal} + \hat{\sigma}^2 \boldsymbol{I}_{N_0+1} \\
&- \tilde{\boldsymbol{X}}_0 [\sum_{i=1}^{I} \tilde{\boldsymbol{X}}_i^{\intercal} \widehat{V_i}^{-1} \tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal} \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}_0^P]^{-1} [\tilde{\boldsymbol{X}}_0^{P\intercal} \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}_0^P \hat{\Sigma} \tilde{\boldsymbol{X}}_0^{\intercal} + \tilde{\boldsymbol{X}}_0^{P\intercal} \widehat{V_0}^{P^{-1}} \\
&\times (\hat{\sigma}^2 \boldsymbol{I}_{N_0}, \boldsymbol{0}_{N_0 \times 1})] - [\tilde{\boldsymbol{X}}_0 \hat{\Sigma} \tilde{\boldsymbol{X}}_0^{P\intercal} \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}_0^P + (\hat{\sigma}^2 \boldsymbol{I}_{N_0}, \boldsymbol{0}_{N_0 \times 1})^{\intercal} \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}_0^P] \\
&\times [\sum_{i=1}^{I} \tilde{\boldsymbol{X}}_i^{\intercal} \widehat{V_i}^{-1} \tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal} \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}_0^P]^{-1} \tilde{\boldsymbol{X}}_0^{\intercal},
\end{aligned}
$$

with

$$
\widehat{V_0}^{FP} := \widehat{\mathbb{C}ov}(Z_0^F, Z_0^P) = \tilde{\boldsymbol{x}}_0^F \hat{\Sigma} \tilde{\boldsymbol{X}}_0^{P\intercal}, \quad \widehat{V_0}^{P} := \widehat{\mathbb{C}ov}(Z_0^P) = \tilde{\boldsymbol{X}}_0^P \hat{\Sigma} \tilde{\boldsymbol{X}}_0^{P\intercal} + \hat{\sigma}^2 \boldsymbol{I}_{N_0},
$$

$$
\tilde{\boldsymbol{X}}_0 := \frac{\partial g_0(a_0, \boldsymbol{x}_0)}{\partial a_0}|_{a_0 = \hat{a}_0}, \quad \tilde{\boldsymbol{X}}_0^P := \frac{\partial g_0^P(a_0, \boldsymbol{x}_0^P)}{\partial a_0}|_{a_0 = \hat{a}_0}, \quad \tilde{\boldsymbol{x}}_0^F := \frac{\partial g(a_0, x_0^F)}{\partial a_0}|_{a_0 = \hat{a}_0},
$$

$$\widehat{\mathbb{C}ov}(\hat{\boldsymbol{\alpha}}) \approx [\sum_{i=1}^{I} \tilde{\boldsymbol{X}}_i^{\intercal} \widehat{V}_i^{-1} \tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal} \widehat{V_0}^{P^{-1}} \tilde{\boldsymbol{X}}_0^{P}]^{-1},$$

where

$$\widehat{V}_i = \tilde{\boldsymbol{X}}_i \hat{\Sigma} \tilde{\boldsymbol{X}}_i^{\intercal} + \hat{\sigma}^2 \boldsymbol{I}_{N_i}, \quad \tilde{\boldsymbol{X}}_i := \frac{\partial g_i(a_i, \boldsymbol{x}_i)}{\partial a_i}|_{a_i=\hat{a}_i}.$$

and $\hat{a}_i$'s are the estimated values of A_i's. Moreover, the $(1-\alpha)$-prediction interval for Y_0^F from a partially observed individual is given by

$$[\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}], \tag{4.51}$$

where $\hat{y}_0^F$ is the realization of $\hat{Y}_0^F$ and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of standard normal distribution.

Proof. The proof can be found in the appendix. □

4.2.5 Prediction approach of Stirnemann et al.

Consider $\hat{\boldsymbol{\beta}} := (\hat{\boldsymbol{\alpha}}, \hat{\sigma}^2, \hat{\Sigma})$ be the estimation of $\boldsymbol{\beta} := (\boldsymbol{\alpha}, \sigma^2, \Sigma)$ which can be obtained based on the information of $\boldsymbol{Y} := (Y_1^{\intercal}, \cdots, Y_I^{\intercal})^{\intercal}$ using one of the proposed estimation procedures for nonlinear random effects model (Pinheiro and Bates, 2000; Walker, 1996; Wang, 2007). Stirnemann et al. (2011) propose estimations of $\mathbb{E}(A_0|y_0^P, \hat{\boldsymbol{\beta}})$ and $\mathbb{E}(Y_0^F|y_0^P, \hat{\boldsymbol{\beta}})$ by using the Markov Chain Monte Carlo (MCMC) algorithm in 2-level nonlinear mixed effects models. They generate the random samples from $f(a_0|y_0^P, \hat{\boldsymbol{\beta}})$, where y_0^P is the realization of Y_0^P, by use of the p-dimensional Metropolis-Hasting algorithm. In the following, we describe the prediction approach of Stirnemann et al. Consider the nonlinear random effects model (4.26) and (4.27) and the nonlinear random effects model for the unknown observation, Y_0^F, (4.28).

Lemma 4.4. *[Stirnemann et al. (2011)]:*
Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimation of $\boldsymbol{\beta}$. For the future observation Y_0^F at x_0^F, the predictive distribution is given by

$$f(y_0^F|y_0^P, \hat{\boldsymbol{\beta}}) = \int f(y_0^F|a_0, \hat{\boldsymbol{\beta}}) f(a_0|y_0^P, \hat{\boldsymbol{\beta}}) da_0.$$

Proof. The proof can be found in the appendix. □

Lemma 4.5. *[Stirnemann et al. (2011)]:*
Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimation of $\boldsymbol{\beta}$ and $(\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)})$ be an M-sample from $f(a_0|y_0^P, \hat{\boldsymbol{\beta}})$ using the Markov Chain Monte Carlo (MCMC) algorithm,

then the Monte Carlo estimations of $\mathbb{E}(Y_0^F|y_0^P,\hat{\boldsymbol{\beta}})$ *and* $\mathbb{E}(A_0|y_0^P,\hat{\boldsymbol{\beta}})$ *are respectively given by*

$$\hat{y}_0^F := \widehat{\mathbb{E}(Y_0^F|y_0^P,\hat{\boldsymbol{\beta}})} \approx \frac{1}{M}\sum_{m=1}^{M} g(\tilde{a}_0^{(m)}, x_0^F), \tag{4.52}$$

$$\hat{a}_0 := \widehat{\mathbb{E}(A_0|y_0^P,\hat{\boldsymbol{\beta}})} \approx \frac{1}{M}\sum_{m=1}^{M} \tilde{a}_0^{(m)}. \tag{4.53}$$

Proof. The proof can be found in the appendix. □

Stirnemann et al. (2011) prove that $\mathbb{E}(Y_0^F|y_0^P,\hat{\boldsymbol{\beta}})$ and $\mathbb{E}(A_0|y_0^P,\hat{\boldsymbol{\beta}})$ are convergent to the conditional expectation of $g(A_0, x_0^F)$ and A_0 given an infinite amount of observations of the new series, respectively. And then based on the simulation results, they show that these conditional expectations and the related true unknown values y_0^F and a_0 are the same, respectively. Hence, it can be concluded that these estimates are unbiased and when more observations of the new series is used, this prediction procedure provide better predictions. Stirnemann et al. (2011) mention that this prediction procedure does not consider uncertainty of the estimated parameter vector $\hat{\boldsymbol{\beta}}$. For taking account this uncertainty, they propose to consider the parameter vector $\boldsymbol{\beta}$ as random with expectation $\hat{\boldsymbol{\beta}}$ and variance covariance matrix equal to the variance of $\hat{\boldsymbol{\beta}}$ and then apply a Bayesian approach.

4.2.5.1 Combination of methods of Pinheiro/Bates and Stirnemann et al. (PBSt)

In this approach, we estimate the parameter vector $\boldsymbol{\beta}$ using the Pinheiro and Bates' estimation method by use of the information of the old series ($\boldsymbol{Y}$). And then by use of the prediction method of Stirnemann et al., we predict the missing observation (Y_0^F). For prediction interval, we use the proposed prediction interval by Stirnemann et al., which is given by

$$PI = [q_{\alpha/2}(\hat{\boldsymbol{\beta}}, y_0^P), q_{1-\alpha/2}(\hat{\boldsymbol{\beta}}, y_0^P)], \tag{4.54}$$

where $q_{\alpha/2}(\hat{\boldsymbol{\beta}}, y_0^P)$ and $q_{1-\alpha/2}(\hat{\boldsymbol{\beta}}, y_0^P)$ are respectively the $(\alpha/2)-$ and $(1-\alpha/2)$-quantiles of $g(\tilde{a}_0^{(1)}, x_0^F), \cdots, g(\tilde{a}_0^{(M)}, x_0^F)$ and $(\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)})$ is an M-sample from $f(a_0|y_0^P,\hat{\boldsymbol{\beta}})$ using the Markov Chain Monte Carlo (MCMC) algorithm. For simulating samples using the MCMC algorithm, the $Metro\_Hastings$ function in the MHadaptive package of Chivers (2012) is used. In this package, a simple Metropolis-Hastings algorithm has been provided with a multivariate normal distribution as a proposal distribution in which the variance-covariance structure is updated at each iteration (Spiegelhalter et al., 2002). For using the $Metro\_Hastings$

function, we only needed to define a function which returns $\log[f(y_0^P|a_0,\hat{\boldsymbol{\beta}})f(a_0|\hat{\boldsymbol{\beta}})]$ and a vector of starting values for a_0 which is obtained from the estimation method of Pinheiro and Bates using the information of $\boldsymbol{Y}_{obs}$.

4.2.5.2 Method of Stirnemann et al. combined with confidence sets (StConf)

In this approach, we use the same estimation and prediction procedures which have been used in Section 4.2.5.1. For prediction interval, we consider the prediction interval based on the confidence ellipsoid for $\boldsymbol{\beta}$. In the following theorem, using the results in Müller et al. (2016) we propose the $(1-\alpha)^2$-prediction interval for Y_0^F based on the $(1-\alpha)$-confidence set for $\boldsymbol{\beta}$.

Theorem 4.4. *If $\hat{\Theta}_\alpha(\boldsymbol{y})$ is the $(1-\alpha)$-confidence set for $\boldsymbol{\beta}$ then the $(1-\alpha)^2$-prediction interval for Y_0^F is given by:*

$$\mathbb{P}(\boldsymbol{y}_{obs}) = \cup_{\boldsymbol{\beta}\in\hat{\Theta}_\alpha(\boldsymbol{y})}[q_{\alpha/2}(\boldsymbol{\beta},y_0^P), q_{1-\alpha/2}(\boldsymbol{\beta},y_0^P)]. \tag{4.55}$$

Proof. The proof can be found in the appendix. □

Let $\hat{\Theta}(\boldsymbol{y})$ be the $(1-\alpha)$-confidence ellipsoid for $\boldsymbol{\beta}$, then based on Theorem 4.4, the $(1-\alpha)^2$-prediction interval for Y_0^F is given by

$$\cup_{\boldsymbol{\beta}\in\hat{\Theta}_\alpha(\boldsymbol{y})}[q_{\alpha/2}(\boldsymbol{\beta},y_0^P), q_{1-\alpha/2}(\boldsymbol{\beta},y_0^P)] \subseteq [\min_{\boldsymbol{\beta}\in\hat{\Theta}(\boldsymbol{y})} q_{\alpha/2}(\boldsymbol{\beta},y_0^P), \max_{\boldsymbol{\beta}\in\hat{\Theta}(\boldsymbol{y})} q_{1-\alpha/2}(\boldsymbol{\beta},y_0^P)],$$

where $q_{\alpha/2}(\boldsymbol{\beta},y_0^P)$ and $q_{1-\alpha/2}(\boldsymbol{\beta},y_0^P)$ are respectively the $(\frac{\alpha}{2})$-and $(1-\frac{\alpha}{2})$-quantiles of $g(\tilde{a}_0^{(1)},x_0^F),\ \cdots, g(\tilde{a}_0^{(M)},x_0^F)$ for each $\boldsymbol{\beta}$ in the confidence ellipsoid $\hat{\Theta}(\boldsymbol{y})$ and $(\tilde{a}_0^{(1)},\cdots,\tilde{a}_0^{(M)})$ is a random sample from $f(a_0|y_0^P,\boldsymbol{\beta})$ (which is obtained from an MCMC algorithm) for each $\boldsymbol{\beta}$ in $\hat{\Theta}(\boldsymbol{y})$.

The confidence ellipsoid for $\boldsymbol{\beta}$ using the Pinheiro and Bates' method can be obtained by use of the *intervals* function in the nlme package of Pinheiro et al. (2016). This confidence ellipsoid is obtained from the asymptotic normality of estimates (see Section 3.2.1.1).

4.2.6 An extension of Walker's estimation method and a modified method of Stirnemann et al.

4.2.6.1 Extending the Walker's estimation method

Consider the nonlinear random effects model (4.26) for the completely observed individuals and (4.27) for the partially observed individual. The two-stage nonlinear

random effects model is given by

$$(Y_i|A_i = a_i) \sim \mathbb{N}_{N_i}(g_i(a_i, \boldsymbol{x}_i), \sigma^2 \boldsymbol{I}_{N_i}), \quad i = 1, \cdots, I,$$

$$(Y_0|A_0 = a_0) \sim \mathbb{N}_{N_0+1}(g_0(a_0, \boldsymbol{x}_0), \sigma^2 \boldsymbol{I}_{N_0+1}),$$

with

$$A_0, \cdots, A_I \overset{iid}{\sim} \mathbb{N}_q(\boldsymbol{\alpha}, \Sigma).$$

Walker (1996) proposes an MCEM algorithm for the estimation of unknown parameters in nonlinear random effects models in which only the random effects (A_i, $i \in \{0, 1, \cdots, I\}$) are considered as the missing data and all individuals have to be completely observed, i.e. there was no missing observation in the individuals. Here we extend the Walker's method by considering the random effects (A_i, $i \in \{0, 1, \cdots, I\}$) and a missing observation in the new individual (Y_0^F) as the missing data in our proposed MCEM algorithm.

MCEM algorithm
Let $\boldsymbol{A} = (A_0, \cdots, A_I), \boldsymbol{Y}_{obs} = (Y_1^\intercal, \cdots, Y_I^\intercal, Y_0^P)^\intercal$ and $\boldsymbol{Z} = (\boldsymbol{Y}_{obs}, Y_0^F, \boldsymbol{A})$ represents the complete data such that $\boldsymbol{A}$ and Y_0^F represent the missing data and also $\boldsymbol{a}, \boldsymbol{y}_{obs}$ and $\boldsymbol{z}$ be the realizations of $\boldsymbol{A}, \boldsymbol{Y}_{obs}$ and $\boldsymbol{Z}$, respectively. Consider $\boldsymbol{\beta} := (\boldsymbol{\alpha}, \Sigma, \sigma^2)$ the parameters vector which should be estimated using the maximum likelihood method. If $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimation of $\boldsymbol{\beta}$ then

$$L(\hat{\boldsymbol{\beta}}|\boldsymbol{y}_{obs}) \geq L(\boldsymbol{\beta}|\boldsymbol{y}_{obs}), \quad \text{for all} \quad \boldsymbol{\beta} \in \Omega,$$

where $\boldsymbol{y}_{obs}$ is the realization of $\boldsymbol{Y}_{obs}$ and Ω is the parameter space.

The E step
At the (k+1)th iteration, define $Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)})$ as follows:

$$Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}) = \mathbb{E}(\log f(\boldsymbol{Z}|\boldsymbol{\beta})|\boldsymbol{y}_{obs}, \boldsymbol{\beta}^{(k)}) = \iint (\log f(\boldsymbol{z}|\boldsymbol{\beta})) f(y_0^F, \boldsymbol{a}|\boldsymbol{y}_{obs}, \boldsymbol{\beta}^{(k)}) dy_0^F d\boldsymbol{a},$$

where

$$\log f(\boldsymbol{z}|\boldsymbol{\beta}) = K - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}(y_i - g_i(a_i, \boldsymbol{x}_i))^\intercal (y_i - g_i(a_i, \boldsymbol{x}_i))$$

$$-\frac{1}{2}\sum_{i=1}^{I}(a_i - \boldsymbol{\alpha})^\intercal \Sigma^{-1}(a_i - \boldsymbol{\alpha}) - \frac{I}{2}\log|\Sigma| \tag{4.56}$$

$$-\frac{1}{2}(a_0-\boldsymbol{\alpha})^\intercal\Sigma^{-1}(a_0-\boldsymbol{\alpha})-\frac{1}{2}\log|\Sigma|-\frac{N_0+1}{2}\log(\sigma^2)$$
$$-\frac{1}{2\sigma^2}\begin{pmatrix} y_0^P-g_0^P(a_0,\boldsymbol{x}_0^P)\\ y_0^F-g(a_0,x_0^F)\end{pmatrix}^\intercal\begin{pmatrix} y_0^P-g_0^P(a_0,\boldsymbol{x}_0^P)\\ y_0^F-g(a_0,x_0^F)\end{pmatrix}, \tag{4.57}$$

for $N:=\sum_{i=1}^I N_i$ and $K:=-(\frac{N+N_0+1+(I+1)q}{2})\log(2\pi)$.
In the Walker's estimation method only the term (4.56) appears. Here we have additionally the term (4.57).

The M step
Let $\hat{\boldsymbol{\beta}}$ be the unique solution at the $(k+1)$th iteration for the following equation:

$$\frac{\partial}{\partial\boldsymbol{\beta}}Q(\boldsymbol{\beta},\boldsymbol{\beta}^{(k)})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}=0. \tag{4.58}$$

By allowing differentiation under the integral sign (Jamshidian and Jennrich, 1993) we have

$$\mathbb{E}\{\frac{\partial}{\partial\boldsymbol{\beta}}\log f(\boldsymbol{Z}|\boldsymbol{\beta})|\boldsymbol{y}_{obs},\boldsymbol{\beta}^{(k)}\}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}=0. \tag{4.59}$$

Then

$$\frac{\partial}{\partial\boldsymbol{\alpha}}\log f(\boldsymbol{z}|\boldsymbol{\beta})=\Sigma^{-1}\sum_{i=0}^I(a_i-\boldsymbol{\alpha})$$

and so

$$\mathbb{E}\{\frac{\partial}{\partial\boldsymbol{\alpha}}\log f(\boldsymbol{Z}|\boldsymbol{\beta})|\boldsymbol{y}_{obs},\boldsymbol{\beta}^{(k)}\}=(\Sigma^{(k)})^{-1}\sum_{i=0}^I\{\mathbb{E}(A_i|y_{i,obs},\boldsymbol{\beta}^{(k)})-\boldsymbol{\alpha}\},$$

where $y_{i,obs}=y_i$, $i\in\{1,\cdots,I\}$ and $y_{0,obs}=y_0^P$ are the realizations of Y_i, $i\in\{1,\cdots,I\}$ and Y_0^P, respectively.
Moreover,

$$\frac{\partial}{\partial\Sigma}\log f(\boldsymbol{z}|\boldsymbol{\beta})=\frac{1}{2}\Sigma^{-1}\{\sum_{i=0}^I[(a_i-\boldsymbol{\alpha})(a_i-\boldsymbol{\alpha})^\intercal-\Sigma]\}\Sigma^{-1},$$

and

$$\frac{\partial}{\partial\sigma^2}\log f(\boldsymbol{z}|\boldsymbol{\beta})=-\frac{N+N_0+1}{2\sigma^2}+\frac{1}{2\sigma^4}[\sum_{i=1}^I(y_i-g_i(a_i,\boldsymbol{x}_i))^\intercal(y_i-g_i(a_i,\boldsymbol{x}_i))$$
$$+\begin{pmatrix} y_0^P-g_0^P(a_0,\boldsymbol{x}_0^P)\\ y_0^F-g(a_0,x_0^F)\end{pmatrix}^\intercal\begin{pmatrix} y_0^P-g_0^P(a_0,\boldsymbol{x}_0^P)\\ y_0^F-g(a_0,x_0^F)\end{pmatrix}].$$

According to the above expressions, the unique solution of (4.59) at the $(k+1)$th iteration is given by

$$\begin{aligned}\hat{\boldsymbol{\alpha}}^{(k+1)} &= \frac{1}{I+1}\sum_{i=0}^{I}\mathbb{E}(A_i|y_{i,obs},\boldsymbol{\beta}^{(k)}) \\ &=: \frac{1}{I+1}\sum_{i=0}^{I}\bar{a}_i\end{aligned} \tag{4.60}$$

$$\begin{aligned}\hat{\Sigma}^{(k+1)} &= \frac{1}{I+1}\sum_{i=0}^{I}\mathbb{E}[(A_i-\hat{\boldsymbol{\alpha}}^{(k+1)})(A_i-\hat{\boldsymbol{\alpha}}^{(k+1)})^{\intercal}|y_{i,obs},\boldsymbol{\beta}^{(k)}] \\ &= \frac{1}{I+1}\sum_{i=0}^{I}[\mathbb{E}(A_iA_i^{\intercal}|y_{i,obs},\boldsymbol{\beta}^{(k)})]-\hat{\boldsymbol{\alpha}}^{(k+1)}\hat{\boldsymbol{\alpha}}^{\intercal(k+1)},\end{aligned} \tag{4.61}$$

and

$$\begin{aligned}\hat{\sigma}^{2(k+1)} &= \frac{1}{N+N_0+1}\{\sum_{i=1}^{I}E[(Y_i-g_i(A_i,\boldsymbol{x}_i))^{\intercal}(Y_i-g_i(A_i,\boldsymbol{x}_i)) \\ &+\begin{pmatrix} Y_0^P-g_0^P(A_0,\boldsymbol{x}_0^P) \\ Y_0^F-g(A_0,x_0^F)\end{pmatrix}^{\intercal}\begin{pmatrix} Y_0^P-g_0^P(A_0,\boldsymbol{x}_0^P) \\ Y_0^F-g(A_0,x_0^F)\end{pmatrix}|\boldsymbol{y}_{obs},\boldsymbol{\beta}^{(k)}]\} \\ &= \frac{1}{N+N_0+1}\{\sum_{i=1}^{I}[(y_i-\mathbb{E}(g_i(A_i,\boldsymbol{x}_i)|y_{i,obs},\boldsymbol{\beta}^{(k)}))^{\intercal}(y_i-\mathbb{E}(g_i(A_i,\boldsymbol{x}_i)|y_{i,obs},\boldsymbol{\beta}^{(k)})) \\ &+\mathrm{tr}(\mathbb{C}\mathrm{ov}(g_i(A_i,\boldsymbol{x}_i)|y_{i,obs},\boldsymbol{\beta}^{(k)}))] \\ &+\begin{pmatrix} y_0^P-\mathbb{E}(g_0^P(A_0,\boldsymbol{x}_0^P)|y_0^P,\boldsymbol{\beta}^{(k)}) \\ \mathbb{E}(Y_0^F-g(A_0,x_0^F)|y_0^P,\boldsymbol{\beta}^{(k)})\end{pmatrix}^{\intercal}\begin{pmatrix} y_0^P-\mathbb{E}(g_0^P(A_0,\boldsymbol{x}_0^P)|y_0^P,\boldsymbol{\beta}^{(k)}) \\ \mathbb{E}(Y_0^F-g(A_0,x_0^F)|y_0^P,\boldsymbol{\beta}^{(k)})\end{pmatrix} \\ &+\mathrm{tr}[\mathbb{C}\mathrm{ov}\{\begin{pmatrix} Y_0^P-g_0^P(A_0,\boldsymbol{x}_0^P) \\ Y_0^F-g(A_0,x_0^F)\end{pmatrix}|y_0^P,\boldsymbol{\beta}^{(k)}\}]\}.\end{aligned} \tag{4.62}$$

The MC step

For the estimation of parameters, the calculation of expectations and covariances in (4.60)-(4.62) with respect to $f(y_0^F,\boldsymbol{a}|\boldsymbol{y}_{obs},\boldsymbol{\beta}^{(k)})$ is done by the Monte Carlo approximation. Consider $K_i := K(A_i)$ an integrable function of A_i $(i=1,\cdots,I)$, then based on the independency between Y_i and Y_0^F and also between Y_0^F and A_i, it is induced that

$$\mathbb{E}(K_i|y_{i,obs},\boldsymbol{\beta}^{(k)}) = \iint k_i f(y_0^F,a_i|y_{i,obs},\boldsymbol{\beta}^{(k)})da_i dy_0^F$$

$$= \iint k_i f(a_i|y_i, \boldsymbol{\beta}^{(k)}) f(y_0^F|y_i, \boldsymbol{\beta}^{(k)}) da_i dy_0^F$$
$$= \int [\int k_i f(a_i|y_i, \boldsymbol{\beta}^{(k)}) da_i] f(y_0^F|\boldsymbol{\beta}^{(k)}) dy_0^F$$
$$= \int k_i f(a_i|y_i, \boldsymbol{\beta}^{(k)}) da_i,$$

where $k_i := K(a_i)$ and

$$f(a_i|y_i, \boldsymbol{\beta}^{(k)}) = \frac{f(y_i|a_i, \boldsymbol{\beta}^{(k)}) f(a_i|\boldsymbol{\beta}^{(k)})}{\int f(y_i|a_i, \boldsymbol{\beta}^{(k)}) f(a_i|\boldsymbol{\beta}^{(k)}) da_i},$$

therefore,

$$\mathbb{E}(K_i|y_{i,obs}, \boldsymbol{\beta}^{(k)}) = \frac{\int k_i f(y_i|a_i, \boldsymbol{\beta}^{(k)}) f(a_i|\boldsymbol{\beta}^{(k)}) da_i}{\int f(y_i|a_i, \boldsymbol{\beta}^{(k)}) f(a_i|\boldsymbol{\beta}^{(k)}) da_i}.$$

Walker (1996) proposes the ratio of Monte Carlo approximations using sampling from $f(a_i|\boldsymbol{\beta}^{(k)})$ as the estimation of $\mathbb{E}(K_i|y_{i,obs}, \boldsymbol{\beta}^{(k)})$ which is given by

$$\widehat{\mathbb{E}(K_i|y_{i,obs}}, \boldsymbol{\beta}^{(k)}) \approx \frac{\sum_{m=1}^{M} K(\tilde{a}_i^{(m)}) f(y_i|\tilde{a}_i^{(m)}, \boldsymbol{\beta}^{(k)})}{\sum_{m=1}^{M} f(y_i|\tilde{a}_i^{(m)}, \boldsymbol{\beta}^{(k)})}, \quad i = 1, \cdots, I,$$

where $\tilde{a}_i^{(1)}, \cdots, \tilde{a}_i^{(M)}$ is a random sample from $f(a_i|\boldsymbol{\beta}^{(k)})$ and $K(\tilde{a}_i^{(m)})$ can be $\tilde{a}_i^{(m)}$, $\tilde{a}_i^{(m)}\tilde{a}_i^{(m)T}$, $g_i(\tilde{a}_i^{(m)}, \boldsymbol{x}_i)$ and $g_i(\tilde{a}_i^{(m)}, \boldsymbol{x}_i) g_i(\tilde{a}_i^{(m)}, \boldsymbol{x}_i)^\intercal$.

For the partially observed individual with one missing observation, we need to generate the random sample from $f(y_0^F, a_0|y_0^P, \boldsymbol{\beta}^{(k)})$. We propose to apply the Gibbs sampler by use of the following full conditional distributions:

$$f(y_0^F|a_0, y_0^P, \boldsymbol{\beta}^{(k)}) = \frac{f(y_0^P, y_0^F, a_0|\boldsymbol{\beta}^{(k)})}{f(y_0^P, a_0|\boldsymbol{\beta}^{(k)})}$$
$$= \frac{f(y_0^P, y_0^F|a_0, \boldsymbol{\beta}^{(k)}) f(a_0|\boldsymbol{\beta}^{(k)})}{f(y_0^P|a_0, \boldsymbol{\beta}^{(k)}) f(a_0|\boldsymbol{\beta}^{(k)})} = \frac{f(y_0^P|a_0, \boldsymbol{\beta}^{(k)}) f(y_0^F|a_0, \boldsymbol{\beta}^{(k)})}{f(y_0^P|a_0, \boldsymbol{\beta}^{(k)})}$$
$$= f(y_0^F|a_0, \boldsymbol{\beta}^{(k)}), \tag{4.63}$$

and

$$f(a_0|y_0^F, y_0^P, \boldsymbol{\beta}^{(k)}) = \frac{f(y_0^P, y_0^F, a_0|\boldsymbol{\beta}^{(k)})}{f(y_0^P, y_0^F|\boldsymbol{\beta}^{(k)})}$$
$$\propto f(y_0^P, y_0^F, a_0|\boldsymbol{\beta}^{(k)}) = f(a_0|\boldsymbol{\beta}^{(k)}) f(y_0|a_0, \boldsymbol{\beta}^{(k)}). \tag{4.64}$$

By the fact that $(Y_0^F|a_0, \boldsymbol{\beta}^{(k)}) \sim \mathbb{N}(g(a_0, x_0^F), \sigma^{2(k)})$ and based on (4.63), the random sample can be generated directly from $(Y_0^F|a_0, \boldsymbol{\beta}^{(k)})$. However, for generating a random sample from $(A_0|y_0^F, y_0^P, \boldsymbol{\beta}^{(k)})$, we propose to apply the Sampling Importance Resampling (SIR) method (Bishop, 2006; Koch, 2007). The considered instrumental density for simulating from (4.64) is $f(a_0|\boldsymbol{\beta}^{(k)})$, hence the importance weights are computed as

$$w_l := \frac{f(a_0^{(l)}|\boldsymbol{\beta}^{(k)}) f(y_0|a_0^{(l)}, \boldsymbol{\beta}^{(k)})}{f(a_0^{(l)}|\boldsymbol{\beta}^{(k)})} = f(y_0|a_0^{(l)}, \boldsymbol{\beta}^{(k)}), \quad l = 1, \cdots, L,$$

where $a_0^{(l)}$ is a random sample from $f(a_0|\boldsymbol{\beta}^{(k)})$ and L should be large enough (see Appendix).

Our proposed Gibbs sampling algorithm along with SIR at the (k+1)th iteration of the EM algorithm is given by

Algorithm 1 Pseudocode for the Gibbs sampler along with the SIR algorithm

1: Draw the initial random sample $\tilde{a}_0^{(0)}$ from $\mathbb{N}_q(\boldsymbol{\alpha}^{(k)}, \Sigma^{(k)})$.
2: **for** m=1 **do** M
3: Simulate $\tilde{y}_0^{F(m)}$ from $\mathbb{N}(g(\tilde{a}_0^{(m-1)}, x_0^F), \sigma^{2(k)})$.
4: **Start of SIR algorithm:**
5: Simulate the L samples $a_0^{(1)}, \cdots, a_0^{(L)}$ from $\mathbb{N}_q(\boldsymbol{\alpha}^{(k)}, \Sigma^{(k)})$.
6: Calculate the importance weights $w_l := f(\begin{pmatrix} y_0^P \\ \tilde{y}_0^{F(m)} \end{pmatrix} |a_0^{(l)}, \boldsymbol{\beta}^{(k)}), \quad l \in \{1, \cdots, L\}$.
7: Calculate the probabilities $p_l := \frac{w_l}{\sum_{l=1}^L w_l}, \quad l \in \{1, \cdots, L\}$.
8: Obtain $\tilde{a}_0^{(m)}$ by taking one sample from $a_0^{(1)}, \cdots, a_0^{(L)}$ with the probabilities $p_1, \cdots, p_L$.
9: **End SIR.**
10: **end for**.

If we consider $(\tilde{y}_0^{F(1)}, \tilde{a}_0^{(1)}), \cdots, (\tilde{y}_0^{F(M)}, \tilde{a}_0^{(M)})$ a random sample from $f(y_0^F, a_0|y_0^P, \boldsymbol{\beta}^{(k)})$ then the expectation estimation of $K(A_0, Y_0^F)$, which is an integrable function of A_0 and Y_0^F, is given by

$$\mathbb{E}(K(A_0, \widehat{Y_0^F})|y_0^P, \boldsymbol{\beta}^{(k)}) \approx \frac{1}{M} \sum_{m=1}^{M} K(\tilde{a}_0^{(m)}, \tilde{y}_0^{F(m)}),$$

where $K(\tilde{a}_0^{(m)}, \tilde{y}_0^{F(m)})$ can be $g_0^P(\tilde{a}_0^{(m)}, \boldsymbol{x}_0^P)$, $(\tilde{y}_0^{F(m)})^2$, $\tilde{y}_0^{F(m)} g(\tilde{a}_0^{(m)}, x_0^F)$. Now the estimation procedure can be started by replacing the estimates of expectations and covariances in (4.60)-(4.62).
Booth and Hobert (1999) propose a termination rule of the MCEM algorithm.

Their suggestion says: when the Monte Carlo EM is implemented it is essential to select the Monte Carlo sample size M at each iteration such that the large Monte Carlo error, which is innate in this method, does not swamp the EM. At the (k+1)th iteration, the EM is said to be swamped by the Monte Carlo error related to $\boldsymbol{\beta}^{(k+1)}$ when the previous value, $\boldsymbol{\beta}^{(k)}$, lies in an approximate $100(1-\alpha)\%$ (e.g., $\alpha = 0.25$) confidence region for $\boldsymbol{\beta}^{(k+1)}$, and in this case M should be increased, e.g., M $\rightarrow M + [M/s]$, where $s \in \{3, 4, 5\}$ (Booth and Hobert, 1999).
In summary, our MCEM algorithm can be presented as follows:

Algorithm 2 Pseudocode for our proposed MCEM algorithm

1: Initialize M, $\boldsymbol{\beta}$.
2: Generate $\tilde{a}_i^{(1)}, \cdots, \tilde{a}_i^{(M)}, \quad i = 1, \cdots, I$ from $\mathbb{N}_q(\boldsymbol{\alpha}^{(k)}, \Sigma^{(k)})$.
3: Generate $(\tilde{a}_0^{(1)}, \tilde{y}_0^{F(1)}), \cdots, (\tilde{a}_0^{(M)}, \tilde{y}_0^{F(M)})$ via algorithm 1.
4: EM step: calculate $\boldsymbol{\beta}^{(k+1)}$ by replacing the associated conditional expectations with their related MC approximations in (4.60)-(4.62).
5: Evaluate the MC error: at the (k+1)th EM iteration, obtain the confidence region of $\boldsymbol{\beta}^{(k+1)}$. If $\boldsymbol{\beta}^{(k)}$ lies in this region, then M $\rightarrow M + [M/s]$, where $s \in \{3, 4, 5\}$.
6: Iterate between 2-5 until convergence.

Wu (2010) mentions that for the Monte Carlo EM algorithm there is no guarantee that likelihood increases at each EM iteration by reason of the Monte Carlo errors in the E step. But, under the regularity conditions the convergency of the MCEM algorithm to a maximum has been proved. Because of the possible reduction of likelihood at each EM iteration, we propose a stopping rule which guarantees the non-decrease of likelihood at each EM iteration. Our suggested stopping rule is: if at the (k+1)th iteration the likelihood value related to $\boldsymbol{\beta}^{(k+1)}$ is smaller than the likelihood value related to $\boldsymbol{\beta}^{(k)}$, we use the former parameter, i.e., $\boldsymbol{\beta}^{(k)}$ as an update parameter value at the (k+1)th iteration, else $\boldsymbol{\beta}^{(k+1)}$ is considered as the update parameter value. We stop the algorithm when reductions happen 200 times for the same parameter.

4.2.6.2 A modified method of Stirnemann et al.

Stirnemann et al. (2011) use the MCMC algorithm to simulate $(\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)})$ from $f(a_0|y_0^P, \hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the ML estimate of $\boldsymbol{\beta}$. In this approach, we propose to simulate $(\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)})$ from $f(a_0|\hat{\boldsymbol{\beta}})$ $((A_0|\hat{\boldsymbol{\beta}}) \sim \mathbb{N}_q(\hat{\boldsymbol{\alpha}}, \hat{\Sigma}))$ and to estimate the conditional expectations using the ratio of two Monte Carlo approximations, which has been proposed by Walker (1996). From proof of Lemma 4.5 we have

$$\mathbb{E}(Y_0^F|y_0^P, \hat{\boldsymbol{\beta}}) = \int g(a_0, x_0^F) f(a_0|y_0^P, \hat{\boldsymbol{\beta}}) da_0 = \frac{\int g(a_0, x_0^F) f(y_0^P|a_0, \hat{\boldsymbol{\beta}}) f(a_0|\hat{\boldsymbol{\beta}}) da_0}{\int f(y_0^P|a_0, \hat{\boldsymbol{\beta}}) f(a_0|\hat{\boldsymbol{\beta}}) da_0},$$

hence

$$\hat{y}_0^F := \mathbb{E}(\widehat{Y_0^F|y_0^P}, \hat{\boldsymbol{\beta}}) \approx \sum_{m=1}^{M} g(\tilde{a}_0^{(m)}, x_0^F) \frac{f(y_0^P|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})}{\sum_{m=1}^{M} f(y_0^P|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})}, \tag{4.65}$$

and

$$\begin{aligned}\mathbb{E}(A_0|y_0^P, \hat{\boldsymbol{\beta}}) &= \int a_0 f(a_0|y_0^P, \hat{\boldsymbol{\beta}}) da_0 \\ &= \frac{\int a_0 f(y_0^P|a_0, \hat{\boldsymbol{\beta}}) f(a_0|\hat{\boldsymbol{\beta}}) da_0}{\int f(y_0^P|a_0, \hat{\boldsymbol{\beta}}) f(a_0|\hat{\boldsymbol{\beta}}) da_0},\end{aligned}$$

therefore,

$$\hat{a}_0 := \mathbb{E}(\widehat{A_0|y_0^P}, \hat{\boldsymbol{\beta}}) \approx \sum_{m=1}^{M} \tilde{a}_0^{(m)} \frac{f(y_0^P|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})}{\sum_{m=1}^{M} f(y_0^P|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})}, \tag{4.66}$$

where $(\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)})$ is a random sample from $f(a_0|\hat{\boldsymbol{\beta}})$ which is obtained from the *mvrnorm* function in the MASS package of Venables and Ripley (2002). Consequently, the estimation of prediction variances can be calculated as follows:

$$\begin{aligned}&\widehat{\mathbb{V}\mathrm{ar}}(Y_0^F|y_0^P, \hat{\boldsymbol{\beta}}) = \mathbb{E}(\widehat{(Y_0^F)^2|y_0^P}, \hat{\boldsymbol{\beta}}) - E(\widehat{Y_0^F|y_0^P}, \hat{\boldsymbol{\beta}})^2 \\ &\approx \hat{\sigma}^2 + \sum_{m=1}^{M} g^2(\tilde{a}_0^{(m)}, x_0^F) \frac{f(y_0^P|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})}{\sum_{m=1}^{M} f(y_0^P|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})} - (\hat{y}_0^F)^2.\end{aligned} \tag{4.67}$$

$$\begin{aligned}&\widehat{\mathbb{C}\mathrm{ov}}(A_0|y_0^P, \hat{\boldsymbol{\beta}}) = \mathbb{E}(\widehat{A_0 A_0^\intercal|y_0^P}, \hat{\boldsymbol{\beta}}) - \mathbb{E}(\widehat{A_0|y_0^P}, \hat{\boldsymbol{\beta}}) \mathbb{E}(\widehat{A_0|y_0^P}, \hat{\boldsymbol{\beta}})^\intercal \\ &\approx \sum_{m=1}^{M} \tilde{a}_0^{(m)} \tilde{a}_0^{(m)\intercal} \frac{f(y_0^P|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})}{\sum_{m=1}^{M} f(y_0^P|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})} - \hat{a}_0 \hat{a}_0^\intercal,\end{aligned} \tag{4.68}$$

Estimation of log-likelihood

The estimation of log-likelihood using the Monte Carlo approximation is given by

$$l(\hat{\boldsymbol{\beta}}) \approx \sum_{i=1}^{I} \log\{\frac{1}{M} \sum_{m=1}^{M} f(y_i|\tilde{a}_i^{(m)}, \hat{\boldsymbol{\beta}})\} + \log\{\frac{1}{M} \sum_{m=1}^{M} f(\hat{y}_0|\tilde{a}_0^{(m)}, \hat{\boldsymbol{\beta}})\},$$

where $\tilde{a}_i^{(1)}, \cdots, \tilde{a}_i^{(M)}$ is a random sample from $f(a_i|\hat{\boldsymbol{\beta}})$ $((A_i|\hat{\boldsymbol{\beta}}) \sim \mathbb{N}_q(\hat{\boldsymbol{\alpha}}, \hat{\Sigma}))$, and $\hat{y}_0 = (y_0^{P\intercal}, \hat{y}_0^F)^\intercal$.

4.2.6.3 Combination of an extended estimation method of Walker and a modified method of Stirnemann et al. (WmSt)

In this approach, we estimate the parameter vector $\boldsymbol{\beta}$ using the extension of Walker's estimation method by use of the information of the old and new series and then predict the missing observation from the partially observed series using the modified method of Stirnemann et al.. For prediction interval, we consider the following prediction interval

$$PI = [q_{\alpha/2}(\hat{\boldsymbol{\beta}}), q_{1-\alpha/2}(\hat{\boldsymbol{\beta}})], \tag{4.69}$$

where $q_{\alpha/2}(\hat{\boldsymbol{\beta}})$ and $q_{1-\alpha/2}(\hat{\boldsymbol{\beta}})$ are respectively the $(\frac{\alpha}{2})$-and $(1-\frac{\alpha}{2})$-quantiles of $g(\tilde{a}_0^{(1)}, x_0^F), \cdots, g(\tilde{a}_0^{(M)}, x_0^F)$ such that $\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)}$ is a random sample from $\mathbb{N}_q(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Sigma}})$.

4.2.6.4 The modified method of Stirnemann et al. combined with confidence sets (mStConf)

In this approach, we use the same estimation and prediction procedures which have been described in Section 4.2.6.3. For prediction interval, we consider the prediction interval based on the confidence sets of parameters. In the following theorem, Nie (2007) presents the confidence sets for the unknown parameters in nonlinear mixed effects models.

Theorem 4.5. *[Nie (2007)]:*
Let $\boldsymbol{\theta} := (\theta_1, \cdots, \theta_w)$ be a vector of unknown structural parameters such that $\Sigma = \Sigma(\boldsymbol{\theta})$ and $N_t := \sum_{i=0}^{I} N_i$ be the total sample size and $\boldsymbol{\theta}_*^{\intercal} := (\boldsymbol{\theta}^{\intercal}, \boldsymbol{\alpha}^{\intercal})$. Consider $\hat{\boldsymbol{\beta}}$ be the ML estimate of $\boldsymbol{\beta}$. Under the conditions 1, 3 and 4 and conditions in Lemma 1 or 2 in Nie (2007), when $N_i, I \to \infty$ it is proved that

$$\sqrt{N_t}(\hat{\sigma}^2 - \sigma^2) \sim \mathbb{N}(0, i_*^{-1}),$$

$$\sqrt{I+1}(\hat{\boldsymbol{\theta}}_* - \boldsymbol{\theta}_*) \sim \mathbb{N}(\mathbf{0}, \Psi^{-1}),$$

where

$$i_* := lim_{N_t \to \infty} \frac{1}{N_t} \sum_{i=0}^{I} \mathbb{E}(-\frac{\partial^2 \log f(Y_{i,obs}, \boldsymbol{\beta})}{\partial(\sigma^2)^2})$$

$$\approx lim_{N_t \to \infty} \frac{1}{N_t} \sum_{i=0}^{I} \mathbb{E}_{Y_{i,obs}}(\tilde{X}_{i\sigma^2}^2(\boldsymbol{\beta}, \boldsymbol{\delta}_i)(\boldsymbol{I}_q + \tilde{Z}_i(\boldsymbol{\beta}, \boldsymbol{\delta}_i)\Sigma\tilde{Z}_i^{\intercal}(\boldsymbol{\beta}, \boldsymbol{\delta}_i))^{-1}),$$

and

$$\Psi := \mathbb{E}[-\frac{\partial^2 \log f(A_i, \boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}_* \partial \boldsymbol{\theta}_*^\intercal}],$$

such that $f(A_i, \boldsymbol{\theta}_*)$ *is the density function of random effects, and if we define* $A_i = \boldsymbol{\delta}_i + \boldsymbol{\alpha}$, $\boldsymbol{\delta}_i \overset{iid}{\sim} \mathbb{N}_q(\mathbf{0}, \Sigma)$ *then we have*

$$lim_{N_i \to \infty} \frac{1}{N_i} \mathbb{E}_{Y_{i,obs}|\boldsymbol{\delta}_i}(-\frac{\partial^2 \log f(Y_{i,obs}|\boldsymbol{\delta}_i, \boldsymbol{\beta})}{(\partial \sigma^2)^2}) \approx \frac{1}{N_i} \tilde{X}^2_{i\sigma^2}(\boldsymbol{\beta}, \boldsymbol{\delta}_i),$$

and

$$\mathbb{E}_{Y_{i,obs}|\boldsymbol{\delta}_i}(-\frac{\partial^2 \log f(Y_{i,obs}|\boldsymbol{\delta}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\delta}_i \partial \boldsymbol{\delta}_i^\intercal}) = \tilde{Z}_i^\intercal(\boldsymbol{\beta}, \boldsymbol{\delta}_i) \tilde{Z}_i(\boldsymbol{\beta}, \boldsymbol{\delta}_i),$$

$Y_{i,obs} = Y_i$ *for the completely observed individuals and* $Y_{0,obs} = Y_0^P$ *for the partially observed individuals.*

In the following theorem, using the results in Müller et al. (2016) we propose the $(1-2\alpha)$-prediction interval based on the $(1-\alpha)$-confidence set.

Theorem 4.6. *Let* Θ *be the parameter space and* $\hat{\Theta}(\boldsymbol{y}_{obs})$ *be the* $(1-\alpha)$*-confidence set, then* $\mathbb{P}(\boldsymbol{y}_{obs})$ *given by*

$$\mathbb{P}(\boldsymbol{y}_{obs}) = \cup_{\boldsymbol{\beta} \in \hat{\Theta}(\boldsymbol{y}_{obs})} [F^{-1}_{\boldsymbol{\beta}}(\alpha/2), F^{-1}_{\boldsymbol{\beta}}(1-\alpha/2)]$$

is the $(1-2\alpha)$*-prediction interval for* Y_0^F*.*

Proof. The proof can be found in the appendix. □

Now, consider $\hat{\Theta}(\boldsymbol{y}_{obs})$ be the $(1-\alpha)$-confidence ellipsoid for $\boldsymbol{\beta}$. Then the $(1-2\alpha)$-prediction interval based on the $(1-\alpha)$ confidence set is given by

$$\cup_{\boldsymbol{\beta} \in \hat{\Theta}(\boldsymbol{y}_{obs})} [q_{\alpha/2}(\boldsymbol{\beta}), q_{1-\alpha/2}(\boldsymbol{\beta})] \subseteq [\min_{\boldsymbol{\beta} \in \hat{\Theta}(\boldsymbol{y}_{obs})} q_{\alpha/2}(\boldsymbol{\beta}), \max_{\boldsymbol{\beta} \in \hat{\Theta}(\boldsymbol{y}_{obs})} q_{1-\alpha/2}(\boldsymbol{\beta})],$$

where $q_{\alpha/2}(\boldsymbol{\beta})$ and $q_{1-\alpha/2}(\boldsymbol{\beta})$ are respectively the $(\frac{\alpha}{2})$-and $(1-\frac{\alpha}{2})$-quantiles of $g(\tilde{a}_0^{(1)}, x_0^F), \cdots, g(\tilde{a}_0^{(M)}, x_0^F)$ for each $\boldsymbol{\beta}$ in the confidence ellipsoid $\hat{\Theta}(\boldsymbol{y}_{obs})$ and $\tilde{a}_0^{(1)}, \cdots, \tilde{a}_0^{(M)}$ is a random sample from $\mathbb{N}_q(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$ for each $(\boldsymbol{\alpha}, \Sigma)$ in $\hat{\Theta}(\boldsymbol{y}_{obs})$.

It should be mentioned that we use different levels of the prediction intervals based on confidence sets in Sections 4.2.5.2 and 4.2.6.4, based on the fact that Y_0^F and $\boldsymbol{Y}$ are independent while Y_0^F and $\boldsymbol{Y}_{obs}$ are dependent.

Confidence sets of parameters:
For this prediction interval based on the confidence sets, we only need to calculate the confidence intervals for the components of Σ and $\boldsymbol{\alpha}$. From Theorem 4.5, we have asymptotically

$$(\hat{\boldsymbol{\theta}}_* - \boldsymbol{\theta}_*) \sim \mathbb{N}_m(\mathbf{0}, [(I+1)\Psi]^{-1}),$$

where $m = q + \frac{q(q+1)}{2}$, q is the dimension of A_i. Let $h_j^\intercal$ be a $(m \times 1)$ vector with a one in the jth position and zeros in all other positions. Then we have

$$h_j(\hat{\boldsymbol{\theta}}_* - \boldsymbol{\theta}_*) \sim \mathbb{N}(0, \frac{1}{I+1} h_j \Psi^{-1} h_j^\intercal),$$

or equivalently

$$(\hat{\boldsymbol{\theta}}_{*j} - \boldsymbol{\theta}_{*j}) \sim \mathbb{N}(0, \frac{1}{I+1} \Psi_{.jj}),$$

where $\Psi_{\cdot} = \Psi^{-1}$. Now we can calculate the confidence interval for each components of $\boldsymbol{\theta}_*$. For getting the $(1-\alpha)$-confidence region, based on the Bonferroni correction we have to calculate the $(1 - \frac{\alpha}{m})$-confidence intervals for each m components of $\boldsymbol{\theta}_*$
.

5 Simulation and Application

5.1 Simulation Results

In this section, a simulation study is applied to compare the obtained MLEs, predictions and prediction intervals from some of the estimation and prediction methods that were discussed previously. We simulated 500 times a data set with 60 individuals, each with 100 observations. These data sets are obtained from the following nonlinear random effects model. For $j = 1, \cdots, 100$ and $i = 1, \cdots, 60$, take

$$E_{ij} \overset{iid}{\sim} \mathbb{N}(0, \sigma^2), \quad A_i \overset{iid}{\sim} \mathbb{N}_3(\boldsymbol{\alpha}, \Sigma),$$

where $\boldsymbol{\alpha} = \begin{pmatrix} 36 \\ -190 \\ -0.7 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 19 & 48 & 0.3 \\ 48 & 846 & 2.5 \\ 0.3 & 2.5 & 0.01 \end{pmatrix}$ and $\sigma^2 = 0.04$ are the estimated parameters from Virkler's data (using all 68 series) if the Pinheiro and Bates' estimation method is used. Then the simulated observed data is given by:

$$Y_{ij} = A_{1i} + A_{2i} x_j^{A_{3i}} + E_{ij},$$

where x_j is the crack length in the Virkler data set.

5.1.1 Applied approaches

In our simulation study, we apply some of the proposed estimation and prediction methods for nonlinear random effects models in Chapters 3 and 4. In the following, a list of the used methods is presented:

WmSt: Combination of an extended estimation method of Walker and a modified method of Stirnemann et al. with the used $M = 5000$ and $L = 100$ (Section 4.2.6.3).

mStConf: The modified method of Stirnemann et al. combined with confidence sets with the used $M = 5000$ and $L = 100$ (Section 4.2.6.4).

PBSt: Combination of methods of Pinheiro/Bates and Stirnemann et al. with the used $M = 5000$ (Section 4.2.5.1).

StConf: Method of Stirnemann et al. combined with confidence sets with the used $M = 5000$ (Section 4.2.5.2).

mPB: The modified method of Pinheiro and Bates (Section 4.2.3).

ePB: Extension of the method of Pinheiro and Bates (Section 3.2.2).

PBLiski: Combination of methods of Pinheiro/Bates and Liski/Nummi (Section 4.2.1).

PBmSwamy: Combination of method of Pinheiro/Bates and the modified method

of Swamy (Section 4.2.2).
Hall: Method of Hall and Clutter (Section 4.2.4).
LR (Simple linear regression in the balanced case): If the explanatory variables x_{ij} do not depend on i and $x_0^F = x_{0N_F}$ then a simple linear regression of Y_{iN_F} on Y_{iN_0} can be used by assuming $Y_{iN_F} = \theta_0 + \theta_1 Y_{iN_0} + Z_i$ for $i = 1, \cdots, I$ where, θ_0, $\theta_1 \in \mathbb{R}$ and Z_i's are errors. The linear regression estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ provide the prediction $\hat{y}_0^F = \hat{y}_{0N_F} = \hat{\theta}_0 + \hat{\theta}_1 y_{0N_0}$ and the prediction interval is given by the classical method for the linear regression (see Section 2.1.2).

For the new individual, we apply the first 10 observations ($N_0 = 10, x_{0N_0} = 10.8$) of the 60th individual (as the new individual) as Y_0^P, and predict the missing observations at $x_0^F = 11$ (as the near future) and $x_0^F = 28.8$ (as the far away future).

5.1.2 Estimation results

In the above mentioned approaches, indeed we use five different estimation methods because the used estimation procedures in methods WmSt and mStConf and in methods StConf, PBSt, ePB and also in methods Hall and mPB are the same. Hence, only the estimation results of methods WmSt, ePB, mPB, PBLiski and PBmSwamy are represented. For the MCEM algorithm of the method WmSt, the Monte Carlo sample size M was considered 5000, and remained constant throughout the iterations. This estimation procedure converges in approximately 400 iterations after 34 hours on average whereas, the estimation procedure of method PBLiski converges in 10 iterations after 20 minutes. The results, including parameter estimates with their biases and MSEs, are represented in Tables 2 and 3 in the appendix. In addition, the impact of the near or far away missing observations on the estimation procedures is studied. The obtained mean squared error (MSE) for the parameter $\beta_l, \quad l \in \{1, \cdots, 10\}$ in Tables 2 and 3 is given by:

$$\mathrm{MSE}_{\beta_l} = \frac{1}{500} \sum_{k=1}^{500} (\beta_l - \hat{\beta}_l^{(k)})^2,$$

where β_l is the true parameter value, and $\hat{\beta}_l^{(k)}$ is the estimation of β_l at the kth simulation. The MSE is a useful criterion for assessing the merit of estimates when the variance is not suitable (because of the bias in the estimates). From Tables 2 and 3, severe bias and large MSE is clear in the estimation of Σ_{11}, Σ_{22} and Σ_{12}, in the methods WmSt, ePB, mPB and PBmSwamy whereas, the method PBLiski generally provides estimations close to the true values with the lowest biases and MSEs for all of the variance covariance components. However, in the estimation of

fixed effects, the methods WmSt, ePB and mPB perform better than the others. In terms of the estimation of error variance (σ^2), all methods relatively provide the same results. Among the five different estimation methods, only in methods WmSt and PBLiski, the predicted future observation is used in the estimation procedure, therefore we can study the effect of near or far away future prediction on these estimates. Based on the results in Tables 2 and 3, the estimation procedure based on the near future prediction generally provides better estimations with the lower biases and MSEs, in comparison with the procedure based on the far away future prediction. Generally, method mPB presents slightly better estimations than method ePB.

5.1.3 Prediction results

In the mentioned approaches in Section 5.1.1, indeed we use the eight different prediction methods because the used prediction procedures in methods WmSt and mStConf and also in StConf and PBSt are the same. Hence, only the prediction results of methods WmSt, PBSt, ePB, mPB, PBLiski, PBmSwamy, Hall and LR are represented. In Figure 2 (Tables 4 and 5 in the appendix), prediction results related to the eight methods have been displayed. The obtained mean squared error (MSE) for the prediction values is given by:

$$\mathrm{MSE}_p = \frac{1}{500}\sum_{k=1}^{500}(Y_0^{F(k)} - \hat{Y}_0^{F(k)})^2,$$

where $Y_0^{F(k)}$ is the real value of Y_0^F at the kth simulation, and $\hat{Y}_0^{F(k)}$ is the prediction of Y_0^F at the kth simulation. From Figure 2, it is clear that the lowest MSE values are related to methods WmSt, PBSt, PBLiski, Hall and LR. Moreover, based on Figure 2 methods ePB and mPB provide predictions with the highest MSE values. It should be noted that the bias and MSE values for the far away future predictions are bigger than those of the near future predictions. It means that the prediction methods perform well for the near future predictions. Method PBmSwamy, for the nearby predictions, performs well however, for the far away future predictions, presents poor results.
Figure 3 (Tables 6 and 7 in the appendix) represents the prediction interval results from the simulation study. In this figure, for assessing performance of the prediction interval, we calculate the interval score of Gneiting and Raftery (2007), which is a combination of the length and the coverage rate of a prediction interval and is given by:

$$S_\alpha^{\mathrm{int}}(l,u,x) = (u-l) + \frac{2}{\alpha}(l-x)\mathbf{1}_{\{x<l\}} + \frac{2}{\alpha}(x-u)\mathbf{1}_{\{x>u\}},$$

where l and u are respectively the lower and upper limits of $(1-\alpha)\times 100\%$ prediction interval and $\mathbf{1}_{\{x<l\}}$ is the indicator function. From Figure 3, the shortest prediction intervals are obtained from methods PBSt, StConf, LR, PBLiski and Hall. Method mStConf gives the largest prediction intervals with the highest coverage rate and interval score among the others. In terms of the length and interval score of prediction intervals, methods PBLiski, Hall, LR and StConf provide better prediction intervals than the others. For the nearby future, method PBmSwamy provides better prediction intervals than for the far away one. Generally, methods ePB and mPB provide relatively the large prediction intervals with the high coverage rate and interval score. Moreover, for the far away future, method WmSt gives better prediction intervals than method PBSt and for the nearby future, method PBSt works better. Method StConf provides the shorter prediction intervals with the lower interval score than method mStConf.

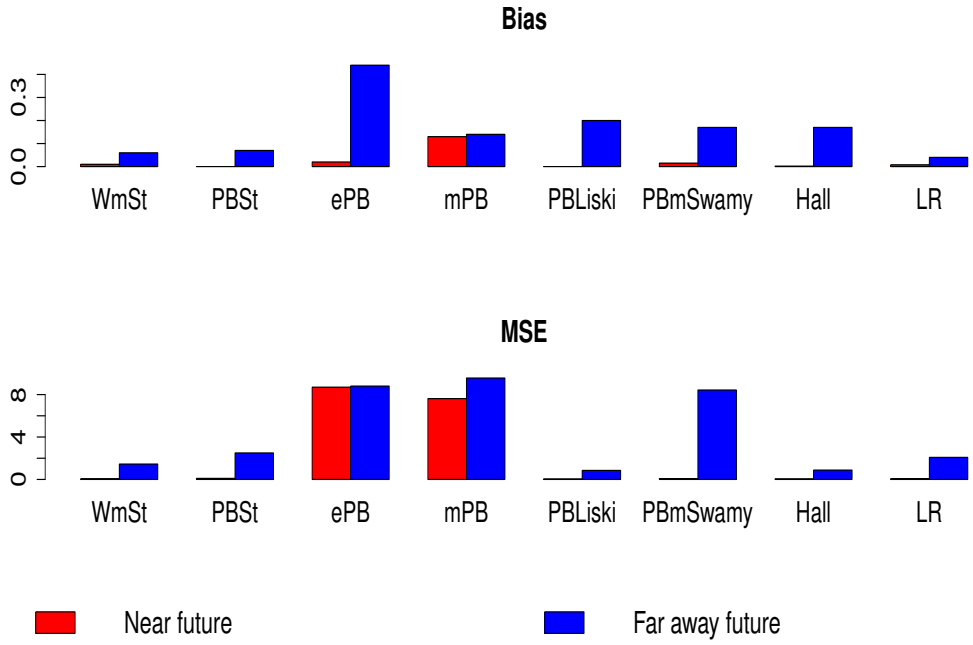

Figure 2: Prediction results from the simulation study at $x_0^F = 11$ (as the near future) and at $x_0^F = 28.8$ (as the far away future)

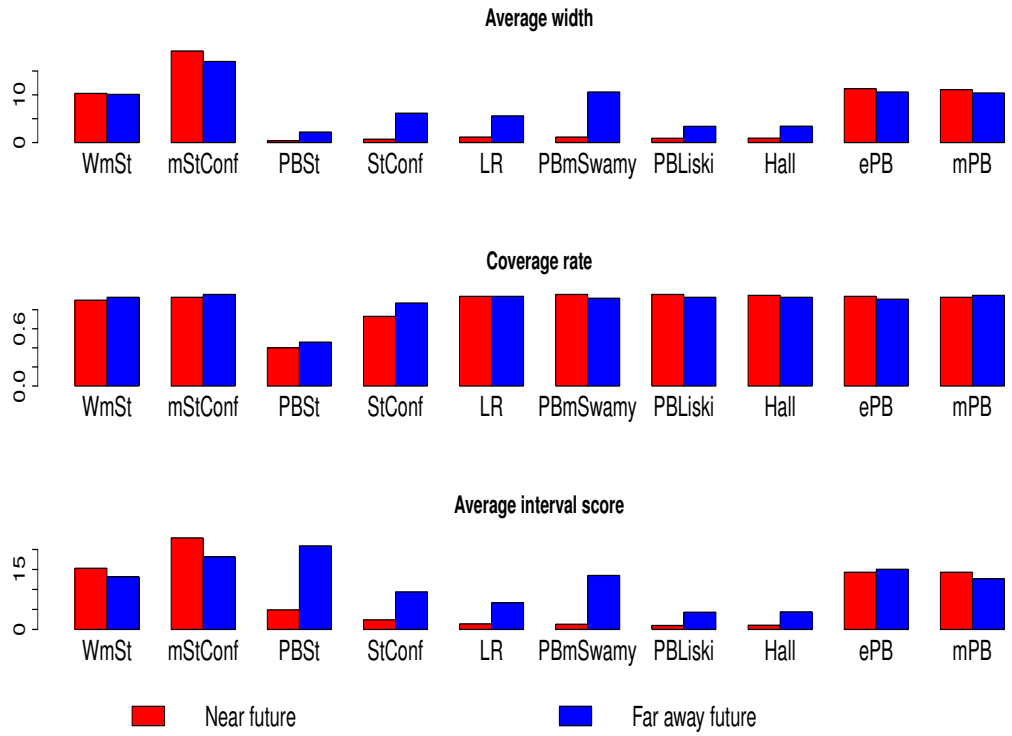

Figure 3: Prediction interval results from the simulation study at $x_0^F = 11$ (as the near future) and at $x_0^F = 28.8$ (as the far away future)

5.2 Application to Virkler's data

The crack growth data presented by Virkler et al. (1979) contains 68 series, where the time was measured when a crack in the steel specimen reached a given length value. The length and time are the explanatory and the dependent variables, respectively. The measurements are at 164 length values, which are the same for all series. The aim is to predict the time, when the crack attains a specific length. For the prediction of future observations, we randomly chose 34 series as the training data set and applied the estimation procedures on them. Then, we predicted some future observations from the other series and supposed that the N_0 observations of them were available. Figure 4 depicts the whole Virkler data set, the red line shows the missing observations which should be predicted.

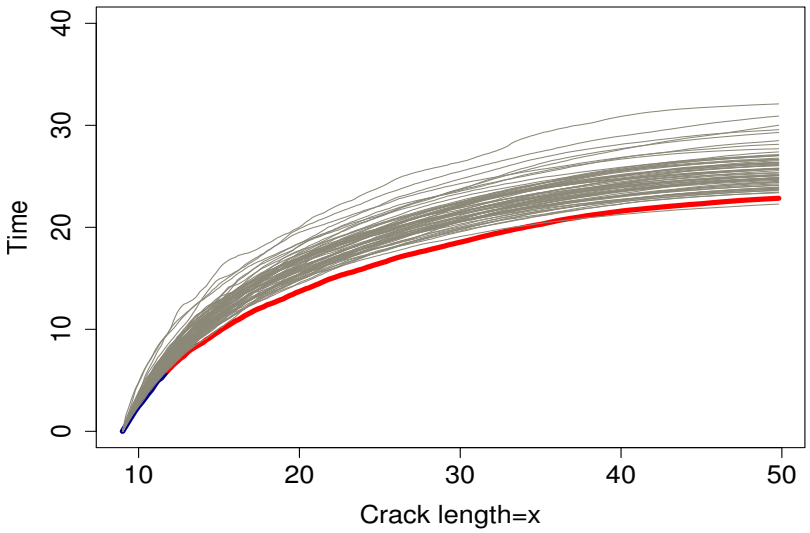

Figure 4: Virkler's data

5.2.1 Applied approaches and models

Two approaches are applied for the Virkler data set: Approach 1) all available observations from the old and new series are applied in the prediction procedures, Approach 2) only the observations from the old series are used in the prediction procedures.

5.2.1.1 Models from the Paris-Erdogan equation

Let l and t be respectively the crack length and the time when the crack attains a specific length. Then, the Paris-Erdogan law for the crack length l at the time t is given by

$$\frac{dl}{dt} = \theta_1 l^{\theta_2}, \tag{5.1}$$

Lemma 5.1. *Depending on θ_2, equation (5.1) has the following solutions:*

(i) $\theta_2 = 1 \Rightarrow l = l(t) = \beta_0 exp(\beta_1 t)$
with $\beta_0 > 0$ *and* $\beta_1 = \theta_1$.
(ii) $\theta_2 < 1 \Rightarrow l = l(t) = \beta_1 (t - \beta_0)^{\beta_2}$
with $\beta_0 < t$, $\beta_1 = (\theta_1(1-\theta_2))^{\frac{1}{1-\theta_2}}$ *and* $\beta_2 = \frac{1}{1-\theta_2} > 0$.
(iii) $\theta_2 > 1 \Rightarrow l = l(t) = \beta_1 (\beta_0 - t)^{-\beta_2}$
with $\beta_0 > t$, $\beta_1 = (\theta_1(\theta_2 - 1))^{-\frac{1}{\theta_2 - 1}}$ *and* $\beta_2 = \frac{1}{\theta_2 - 1} > 0$.

Since in Virkler's data, length l is independent variable and time t is dependent variable hence we need to calculate the inverse $t(l)$ of the functions $l(t)$ of Lemma 5.1.

Lemma 5.2. *The inverse functions $t(l)$ of the functions $l(t)$ of Lemma 5.1 are given by:*

(i) $\theta_2 = 1 \Rightarrow t = t(l) = \alpha_0 + \alpha_1 \log(l)$.
with $\alpha_0 = -\frac{1}{\beta_1}\log(\beta_0)$, $\alpha_1 = \frac{1}{\beta_1}$.
(ii) $\theta_2 < 1 \Rightarrow t = t(l) = \alpha_0 + \alpha_1 l^{\alpha_2}$.
with $\alpha_0 = \beta_0 < t$, $\alpha_1 = (\frac{1}{\beta_1})^{\frac{1}{\beta_2}} > 0$, $\alpha_2 = \frac{1}{\beta_2} > 0$.
(iii) $\theta_2 > 1 \Rightarrow t = t(l) = \alpha_0 + \alpha_1 l^{\alpha_2}$.
with $\alpha_0 = \beta_0 > t$, $\alpha_1 = -(\frac{1}{\beta_1})^{-\frac{1}{\beta_2}} < 0$, $\alpha_2 = -\frac{1}{\beta_2} < 0$.

All obtained models from Lemma 5.1 and Lemma 5.2 are linear and nonlinear in the unknown parameters. Because crack growth is not deterministic, we can obtain a simple stochastic version of these functions by adding a random error. This can create the following models:
1. Linear model: $Y_{ij} = \alpha_0 + \alpha_1 \log(x_{ij}) + E_{ij}$,
2. Nonlinear model: $Y_{ij} = \alpha_0 + \alpha_1 x_{ij}^{\alpha_2} + E_{ij}$,
3. Linearized model of the nonlinear model: $Y_{ij} = \alpha_0 + \alpha_1 x_{ij} + \alpha_2 x_{ij} \log(x_{ij}) + E_{ij}$,
where Y_{ij} is the time variable and x_{ij}'s are the given length values and E_{ij} is the random error.
As it was mentioned in Chapter 1, we apply random effects models, i.e. the parameters α_0, α_1, α_2 are replaced by the individual parameters a_{0i}, a_{1i}, a_{2i} which

are the realizations of random variables, A_{0i}, A_{1i}, A_{2i}. In the following, the three obtained models from the Paris-Erdogan equation , which are used in the two mentioned approaches, are presented:

1, **2**. The linear random effects models, which are modeled as:

$$Y_{ij} = \boldsymbol{x}_j^{\intercal} A_i + E_{ij}, \quad i = 1, \cdots, 34, \quad j = 1, \cdots, 164,$$

the new individual is modeled as:

$$Y_{0j}^P = \boldsymbol{x}_j^{\intercal} A_0 + E_{0j}, \quad j = 1, \cdots N_0,$$

where $\boldsymbol{x}_j := (1, \log(x_j))^{\intercal}$ for the linear model L2 (linear model with two parameters), and $\boldsymbol{x}_j := (1, x_j, x_j \log(x_j))^{\intercal}$ for the linearized model L3 (linear model with three parameters), and for $i = 1, \cdots, 34$ and $j = 1, \cdots, 164$ we have $E_{ij} \sim \mathbb{N}(0, \sigma^2)$, $A_i \sim \mathbb{N}_2(\boldsymbol{\alpha}, \Sigma)$ for the linear model L2 and $A_i \sim \mathbb{N}_3(\boldsymbol{\alpha}, \Sigma)$ for the linear model L3. For $i = 0$ and $j = 1, \cdots, N_0$, we have $E_{0j} \sim \mathbb{N}(0, \sigma^2)$, $A_0 \sim \mathbb{N}_2(\boldsymbol{\alpha}, \Sigma)$ for the linear model L2 and $A_0 \sim \mathbb{N}_3(\boldsymbol{\alpha}, \Sigma)$ for the linear model L3.
It is also supposed that $A_0, A_1, \cdots, A_{34}, E_0^P, E_1, \cdots, E_{34}$ are independent, where $E_i := (E_{i1}, \cdots, E_{i164})^{\intercal}$ and $E_0^P := (E_{01}, \cdots, E_{0N_0})^{\intercal}$.

3. The nonlinear random effects model, which is modeled as

$$Y_{ij} = g(A_i, x_j) + E_{ij}, \quad i = 1, \cdots, 34, \quad j = 1, \cdots, 164,$$

the new individual is modeled as:

$$Y_{0j}^P = g(A_0, x_j) + E_{0j}, \quad j = 1, \cdots, N_0,$$

where for $i = 1, \cdots, 34$ and $j = 1, \cdots, 164$ we have:

$$g(A_i, x_j) := A_{1i} + A_{2i} x_j^{A_{3i}}, \quad E_{ij} \sim \mathbb{N}(0, \sigma^2),$$

$$A_i := (A_{1i}, A_{2i}, A_{3i})^{\intercal} \sim \mathbb{N}_3(\boldsymbol{\alpha}, \Sigma),$$

and for $i = 0$ and $j = 1, \cdots, N_0$ we have:

$$g(A_0, x_j) := A_{10} + A_{20} x_j^{A_{30}}, \quad E_{0j} \sim \mathbb{N}(0, \sigma^2),$$

$$A_0 := (A_{10}, A_{20}, A_{30})^{\intercal} \sim \mathbb{N}_3(\boldsymbol{\alpha}, \Sigma).$$

It is also supposed that $A_0, A_1, \cdots, A_{34}, E_0^P, E_1, \cdots, E_{34}$ are independent, where $E_i := (E_{i1}, \cdots, E_{i164})^{\intercal}$ and $E_0^P := (E_{01}, \cdots, E_{0N_0})^{\intercal}$.

In the following, all discussed prediction methods and models are listed:

Approach 1 (Using all observations from the old and new series):
WmSt: Combination of an extended estimation method of Walker and a modified method of Stirnemann et al. with the used $M = 5000$ and $L = 100$ (Section 4.2.6.3).
mStConf: The modified method of Stirnemann et al. combined with confidence sets with the used $M = 5000$ and $L = 100$ (Section 4.2.6.4).
PBSt: Combination of methods of Pinheiro/Bates and Stirnemann et al. with the used $M = 5000$ (Section 4.2.5.1).
StConf: Method of Stirnemann et al. combined with confidence sets with the used $M = 5000$ (Section 4.2.5.2).
mPB: The modified method of Pinheiro and Bates (Section 4.2.3).
PBLiski: Combination of methods of Pinheiro/Bates and Liski/Nummi (Section 4.2.1).
PBmSwamy: Combination of method of Pinheiro/Bates and the modified method of Swamy (Section 4.2.2).
Hall: Method of Hall and Clutter (Section 4.2.4).
mRao.L2, mRao.L3: The modified method of Rao for the linear models L2 and L3 (Section 4.1.1).
eRao.L2, eRao.L3: Extension of the method of Rao for the linear models L2 and L3 (Section 4.1.1.1).
LR: Simple linear regression in the balanced case (Section 5.1.1).

Approach 2 (Using only observations from the old series):
ePB: Extension of the method of Pinheiro and Bates (Section 3.2.2).
Swamy.L2, Swamy.L3: Method of Swamy for the linear models L2 and L3 (Section 3.1.2).

5.2.2 Estimation results

For the estimation of parameters in the Virkler data set, we chose randomly 34 series as the training data set and obtained the estimations and corresponding standard errors based on the information of these 34 series and the beginning of a new one, which is one of the other 34 series. Required starting values for the estimation procedures are obtained based on the information of the training data set and the new series. Because based on the simulation results, the estimation results related to the near future prediction in methods WmSt and PBLiski are slightly better than the results related to the far away future prediction hence,

in Table 8 we only represented the estimation results related to the near future prediction using the first 131 observations of the new series. The estimations and their related standard errors, in Table 8, are obtained as follows:

$$\hat{\beta}_l = \frac{1}{34}\sum_{k=1}^{34}\hat{\beta}_l^{(k)}, \quad \hat{\sigma}_{\hat{\beta}_l} = \frac{1}{34}\sum_{k=1}^{34}\hat{\sigma}_{\hat{\beta}_l^{(k)}}$$

where $\hat{\beta}_l^{(k)}$ and $\hat{\sigma}_{\hat{\beta}_l^{(k)}}$ are the obtained estimation and its standard error of parameter β_l, $l \in \{1, \cdots, 10\}$ based on the 34 series in the training data set and the beginning of the kth new series. The extension of Walker's estimation method, which is used in methods WmSt and mStConf, converged on average in 300 iterations after 48 hours. The value of M was considered 5000, and remained constant throughout the iterations. The estimation procedure in method PBLiski converged on average in 5 iterations after about 10 minutes. It can be seen that all four methods result in the close parameter estimates except for the estimates of Σ_{11}, Σ_{12} and Σ_{22}. Based on the simulation results, method PBLiski provides the nearest estimates of variance covariance components to the actual ones, and methods WmSt and mPB present better estimates of the fixed effects. With respect to Table 8, it is clear that differences between the related estimated parameter values (especially for Σ_{11}, Σ_{12} and Σ_{22}) to the four methods in the Virkler data set are smaller than those of the simulation study. In other words, methods WmSt, mPB and PBmSwamy perform as well as method PBLiski when the number of given observations within each individual increases (In the Virkler data set, the number of intraindividual is 164 whereas in the simulation study was 100).

5.2.3 Prediction results

For the prediction of future observations in the Virkler data set, 34 series were randomly chosen as the training data set. The prediction is performed by using the other 34 series. We studied the performance of the prediction methods with two sizes of observations from the new series. We initially considered using 15 observations, i.e., up to the crack length of 11.8, and then used 131 observations, i.e., up to the crack length of 35. For the linear regression method, we first considered $x_{0N_0} = 11.8$ and $x_0^F = (12, 17, 24, 29, 33, 36, 40.2, 43, 45, 49.8)$, and finally $x_{0N_0} = 35$ and $x_0^F = (35.2, 36, 37, 38.2, 39, 40.2, 41, 43, 45, 49.8)$. In methods PBSt, StConf, WmSt and mStConf, the Monte Carlo sample size for prediction of the future observation was 5000. Moreover, we considered 35 and 50 points inside the confidence sets of parameters for calculating the prediction intervals in methods StConf and mStConf, respectively. In the following, the obtained results are presented.

5.2.3.1 Prediction results for $x >= 12$:

Prediction results related to the linear models:
Based on Figure 5, method LR provides predictions with the lowest MSE values. Method eRao.L3 presents better predictions with the lower bias and MSE values than method eRao.L2. Methods mRao.L2 and mRao.L3 provides the worst prediction results, especially for the far away future predictions. Generally, Swamy's prediction method perform better than methods eRao and mRao.
From Figure 6, the shortest prediction interval with the high coveragte rate and the lowest interval score is given by method LR. The Swamy's prediction gives relatively large prediction intervals with the highest coverage rate and low interval score. For the near future, method mRao.L2 presents better prediction intevals than method eRao.L2 and for the far away future, method eRao.L2 works better. Generally, methods eRao.L3 and mRao.L3 give better results than eRao.L2 and mRao.L2 and the modified method of Rao results in the larger prediction intervals than the method of eRao.

Prediction results related to the nonlinear model:
In the mentioned approaches in Section 5.2.1, indeed we use the eleven different prediction methods because the used prediction procedures in methods WmSt and mStConf and also in StConf and PBSt are the same. Hence, for the nonlinear model only the prediction results of methods WmSt, PBSt, ePB, mPB, PBLiski, PBmSwamy, Hall and LR are represented. Based on Figure 7, method PBmSwamy gives predictions with the high bias and MSE values among the othe other methods. Method LR provides predictions with the lowest MSEs but with the high bias values. Methods PBLiski and Hall give the similar prediction results with relatively the low MSE and bias values.
From Figure 9, method mPB gives predictions with the lower biases and MSEs especially, for the far away future predictions in comparison to method ePB. Based on Figure 11, for the nearby future, the prediction approach of Stirnemann et al. provides better predictions with the lower biases and MSEs in comparison to its modification method and for the far away future, the modified method of Stirnemann et al. works better.
From Figure 8, the shortest prediction intervals with the low coverage rate and high interval score are obtained from methods Hall and PBLiski. Methods mStConf and StConf give the large prediction intervals with the highest coverage rate and the lowest interval score. Method WmSt provides the larger prediction intervals with the higher coverage rate and the lower interval score than method PBSt. Method LR gives the short prediction intervals with the high coveragre rate and the low interval score. Large prediction intervals with high coverage rate and low

interval score are obtained from methods PBmSwamy, ePB and mPB.
Based on Figure 10, method mPB presents prediction intervals with the smaller interval score than method ePB. With respect to Figure 12, the best prediction intervals with the lowest interval score and the highest coverate rate are given by methods mStConf and StConf. Moreover, method WmSt provides better prediction intervals than method PBSt.

An overall comparison:
For an overall comparison of methods for linear and nonlinear models, we chose methods which provide the higher coverage rate or the lower interval score than the other methods. Hence from Figure 6, we considered methods Swamy.L2, Swamy.L3 and eRao.L3 and from Figure 8, methods mStConf, StConf, LR, mPB, WmSt and PBmSwamy were selected. Based on Figure 13, methods LR and mPB give predictions with the lowest MSEs. From Figure 14, in terms of coverage rate and interval score of the prediction intervals, methods mStConf, StConf and mPB present the best prediction intervals.

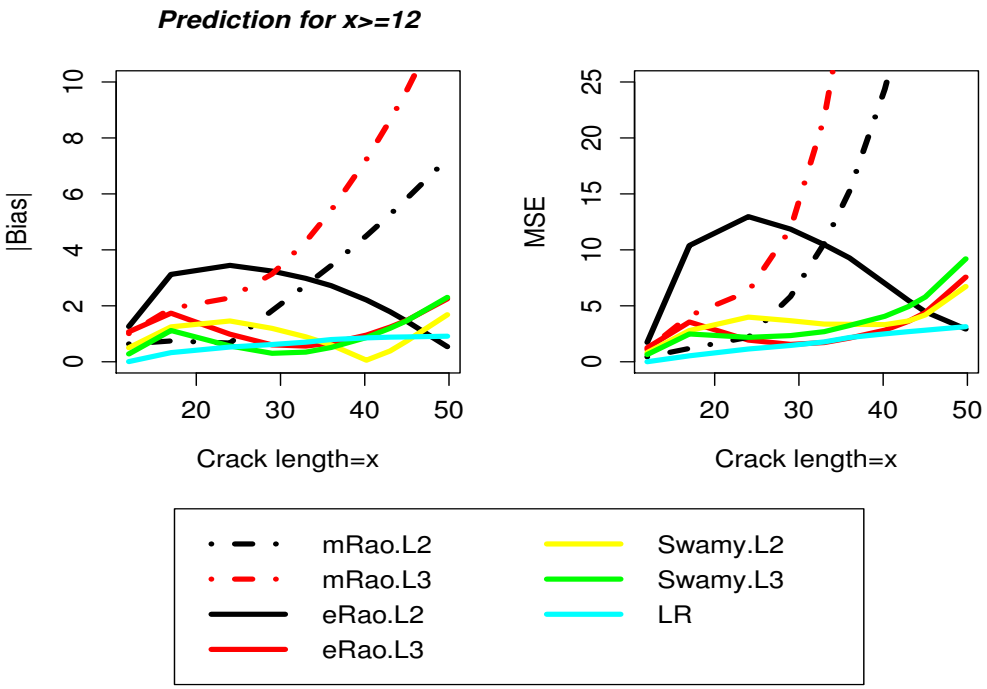

Figure 5: Prediction results of the linear models L2 and L3 using the first 15 observations of the new series

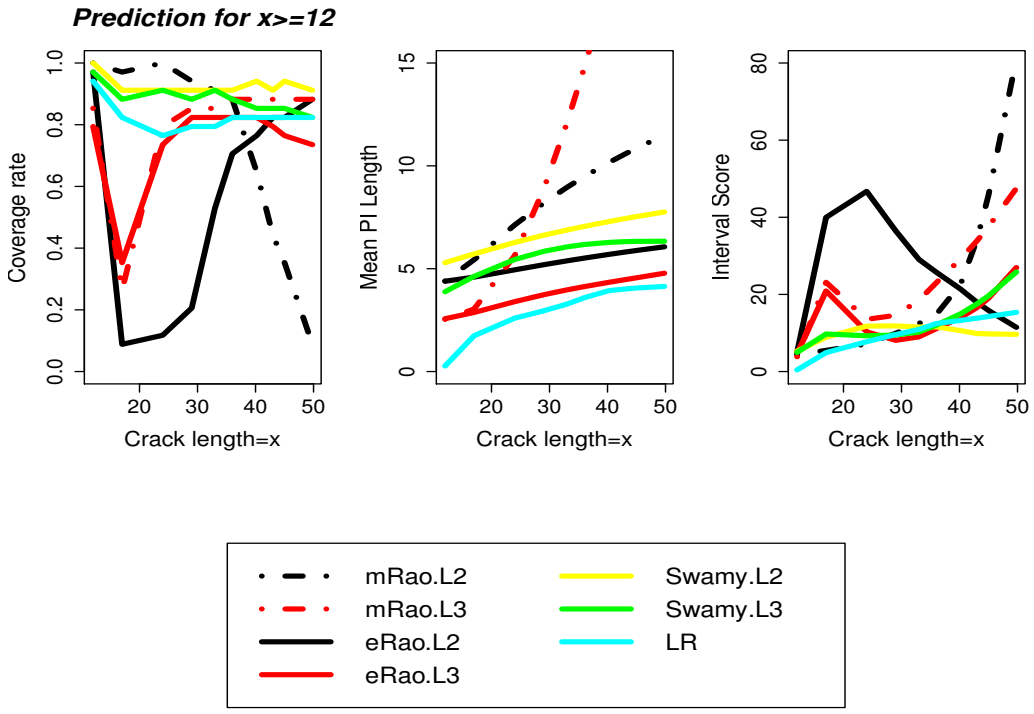

Figure 6: Prediction intervals of the linear models L2 and L3 using the first 15 observations of the new series

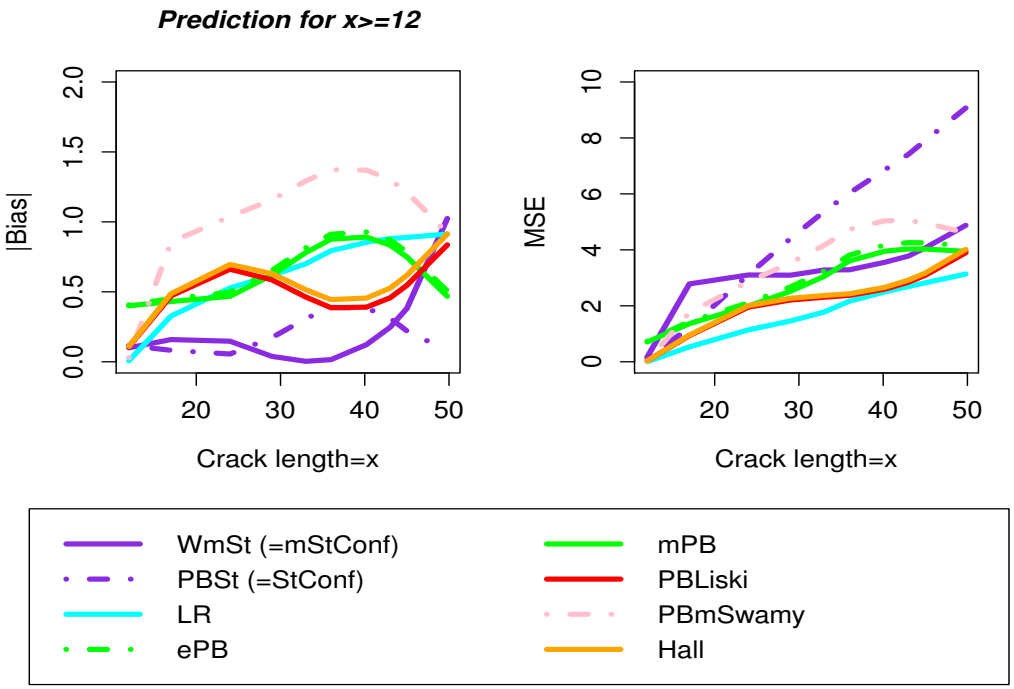

Figure 7: Prediction results of the nonlinear model using the first 15 observations of the new series

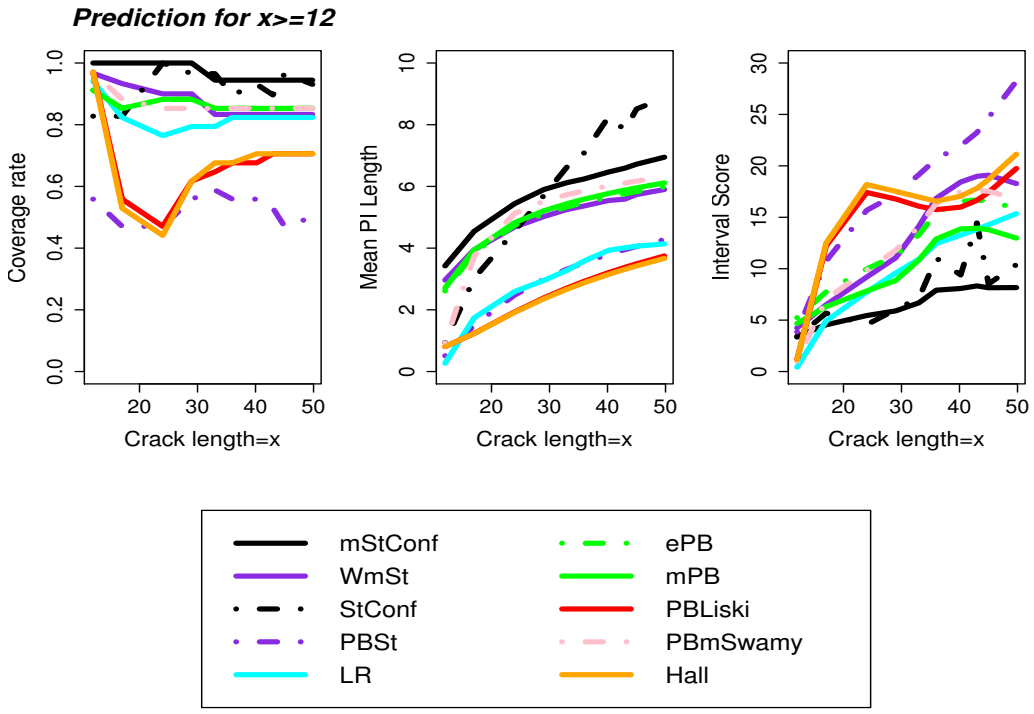

Figure 8: Prediction intervals of the nonlinear model using the first 15 observations of the new series

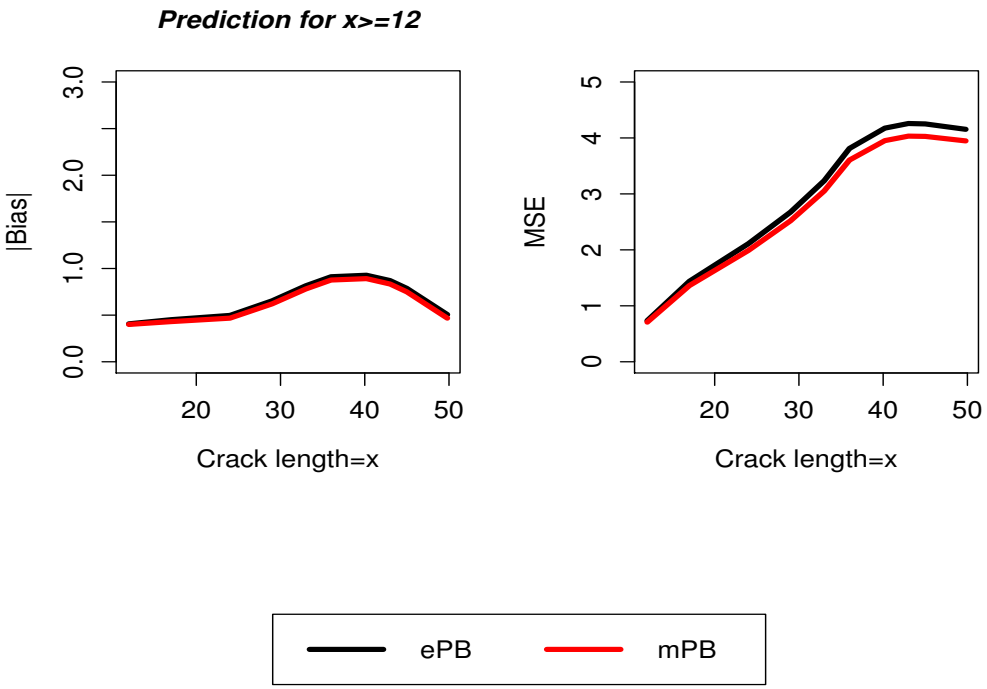

Figure 9: Prediction results of methods ePB and mPB using the first 15 observations of the new series

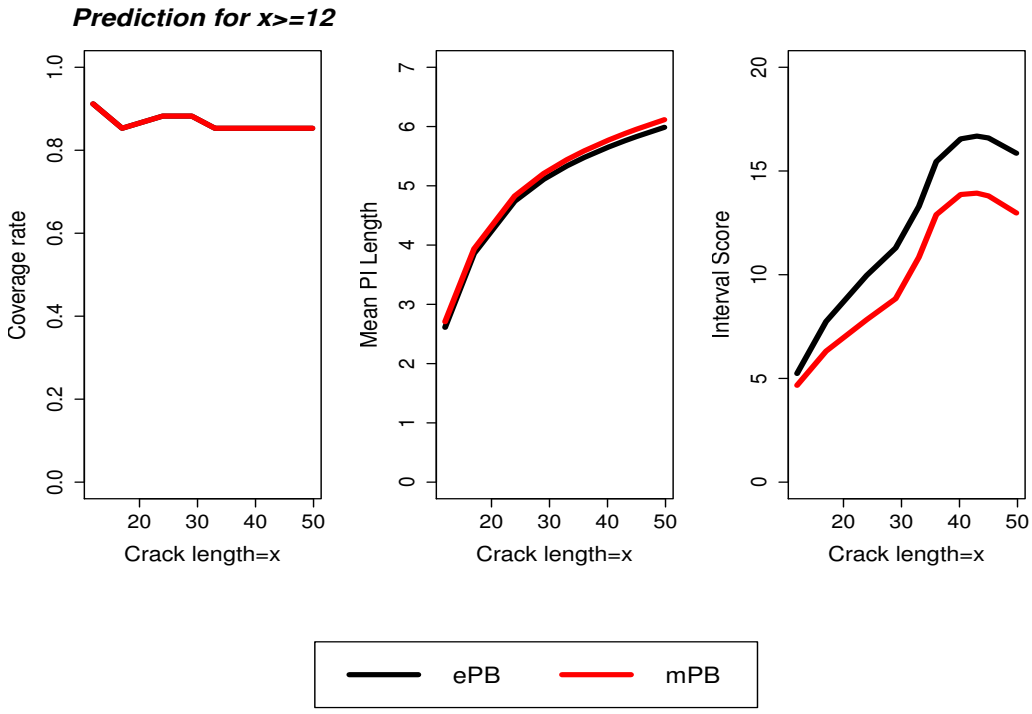

Figure 10: Prediction intervals of methods ePB and mPB using the first 15 observations of the new series

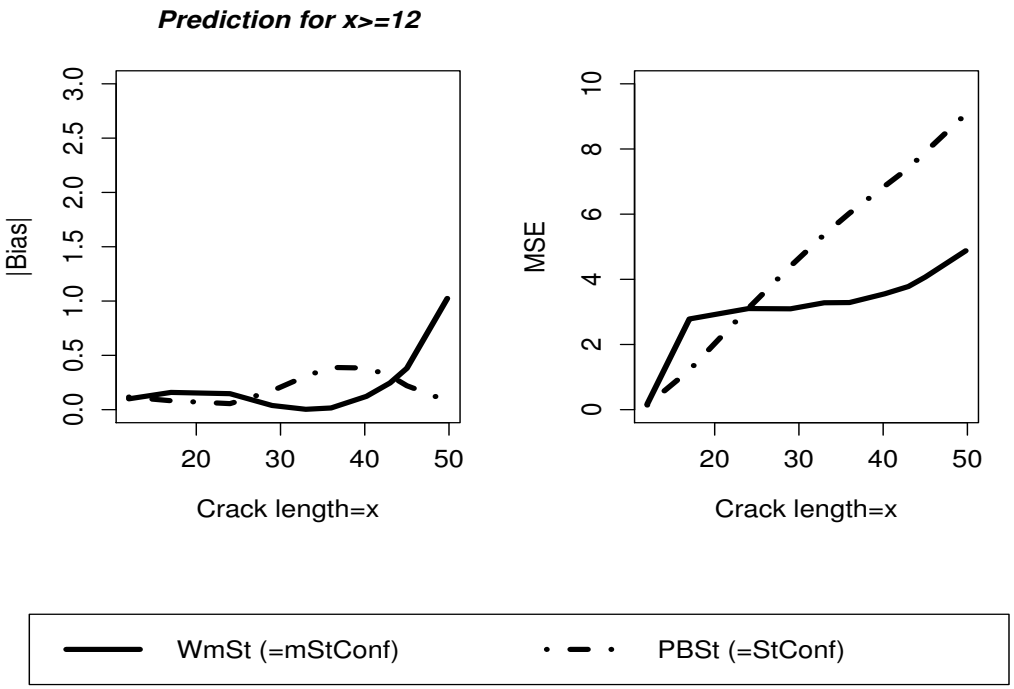

Figure 11: Prediction results of methods WmSt and PBSt using the first 15 observations of the new series

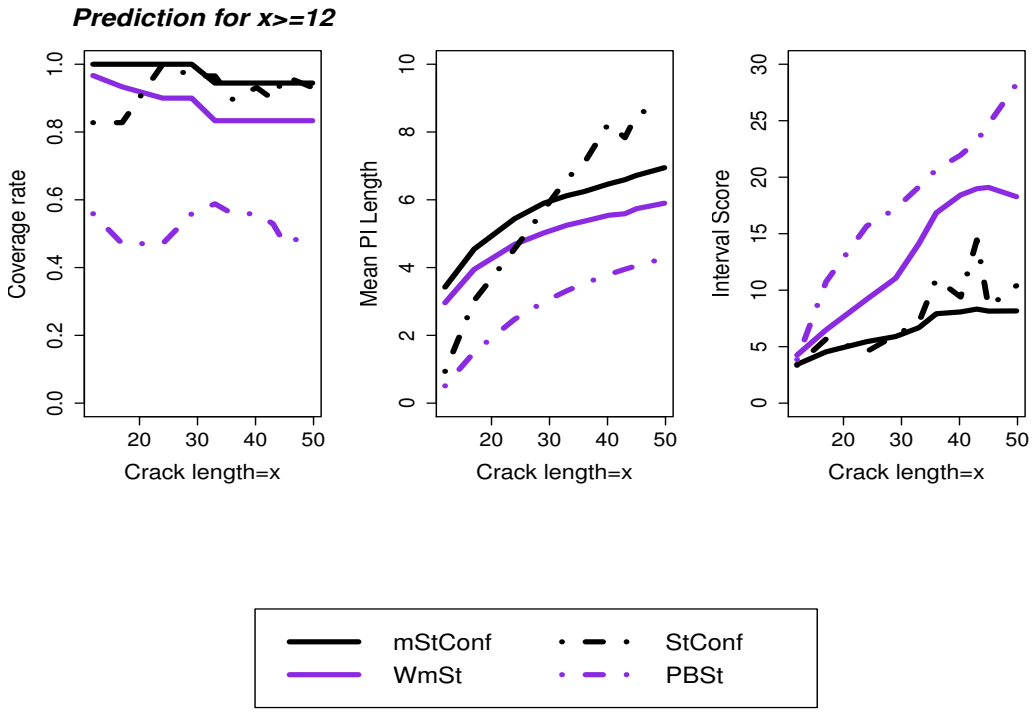

Figure 12: Prediction intervals of methods WmSt, PBSt, StConf and mStConf using the first 15 observations of the new series

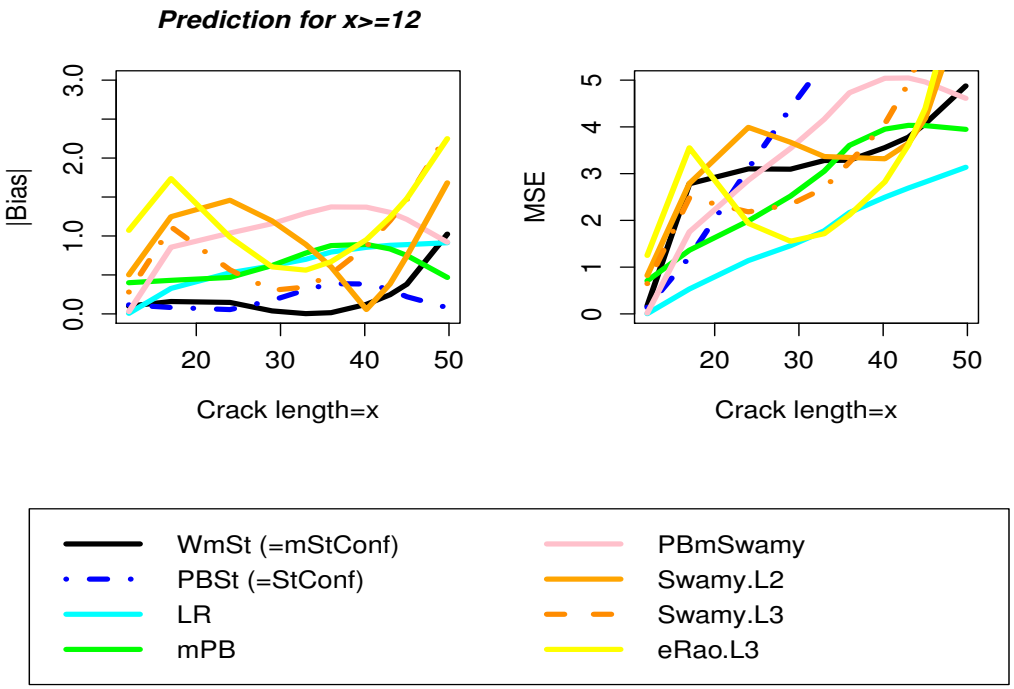

Figure 13: Overall comparison of prediction results for $x >= 12$

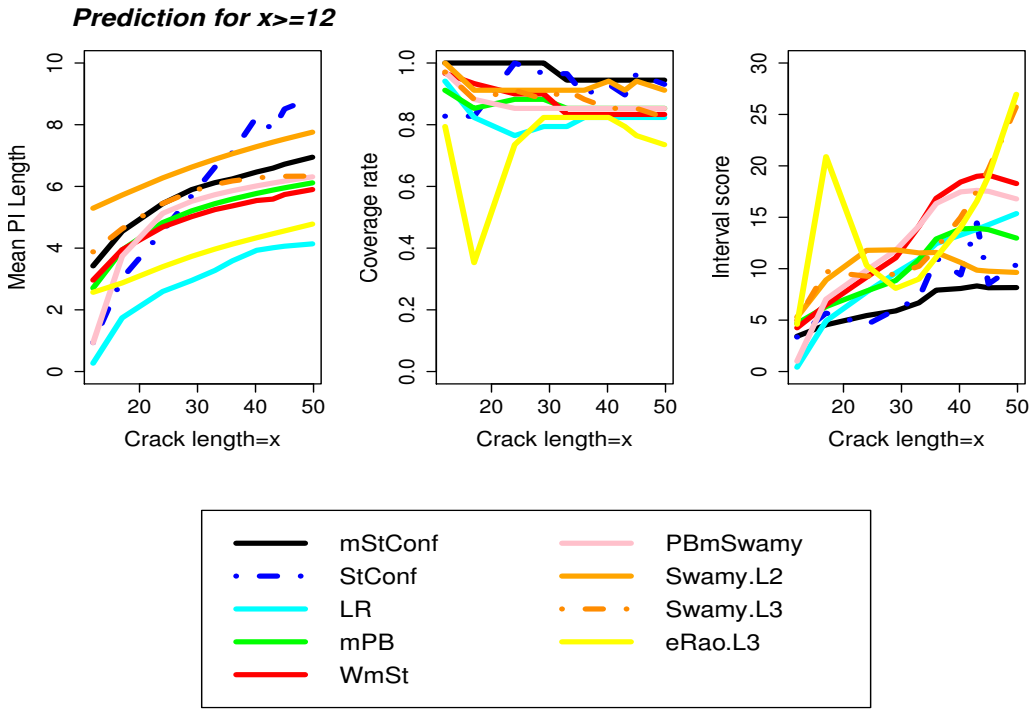

Figure 14: Overall comparison of prediction intervals for $x >= 12$

5.2.3.2 Prediction results for $x > 35$:

Prediction results related to the linear models:
From Figure 15, method LR gives the best predictions with the smallest bias and MSE values. Methods eRao.L3, mRao.L3, Swamy.L2 and Swamy.L3 provide predictions with the low bias and MSE values. Methods eRao.L2 and mRao.L2 give the worst predictions with the highest bias and MSE values especially, for the far away future predictions.
Based on Figure 16, the shortest prediction intervals with the high coverage rate and the smallest interval score is given by method LR. Methods Swamy.L2 and Swamy.L3 give the largest prediction intervals with the highest coverage rate and small interval score. Methods eRao.L3 and mRao.L3 give the shorter prediction intervals with the higher coverage rate and the lower interval score than methods eRao.L2 and mRao.L2.

Prediction results related to the nonlinear model:
In the mentioned approaches in Section 5.2.1, indeed we use the eleven different prediction methods because the used prediction procedures in methods WmSt and mStConf and also in StConf and PBSt are the same. Hence, for the nonlinear model only the prediction results of methods WmSt, PBSt, ePB, mPB, PBLiski, PBmSwamy, Hall and LR are represented. From Figure 17, the best predictions with the smallest bias and MSE values are given by method LR. Methods PBSt, PBLiski, PBmSwamy and Hall provide predictions with the low bias and MSE values. Methods ePB and mPB provide the worst predictions with the highest bias and MSE values. Method WmSt gives predictions with the high bias and MSE values especially for the far away future predictions.
From Figure 19, method mPB provides better predictions with the lower bias and MSE values than method ePB. In addition, based on Figure 21, method PBSt provides better predictions with the lower bias and MSE values than method WmSt.
From Figure 18, the shortest prediction intervals with the high coverage rate and the smallest interval score are given by method LR. The largest prediction intervlas with the highest coverage rate and the high interval score is obtained from method mStConf. Methods PBSt, PBLiski, PBmSwamy, LR and Hall give the shortest prediction intervals. Method StConf provides the short prediction intervals with the highest coverage rate and the low interval score. The linearization-based methods (i.e., PBLiski, PBmSwamy and Hall) provide the shortest prediction intervals with the relatively high coverage rate and small interval score. Methods mStConf, WmSt, ePB and mPB give the largest prediction intervals with the highest coverage rate and the highest interval score.
From Figure 20, method mPB gives better prediction intervals with the higher

coverage rate and the lower interval score than method ePB. From Figure 22, methods mStConf and WmSt give the largest prediction intervals with the high coverage rate and the highest interval score. Method StConf provides the best prediction intervals with the highest coverage rate and the lowest interval score.

An overall comparison:
For an overall comparison of methods for linear and nonlinear models, we chose methods which provide the higher coverage rate or the lower interval score than the other methods. Hence from Figure 16 we considered methods Swamy.L2 and Swamy.L3 and from Figure 18, methods mStConf, StConf, WmSt, LR, mPB, PBLiski, PBmSwamy and Hall were selected. Based on Figures 23, the best predictions with the smallest MSEs are given by methods LR, Hall, PBSt, PBLiski and PBmSwamy. From Figure 24, the best prediction intervals with the lowest interval score are given by methods LR, StConf, Hall, PBLiski and PBmSwamy.

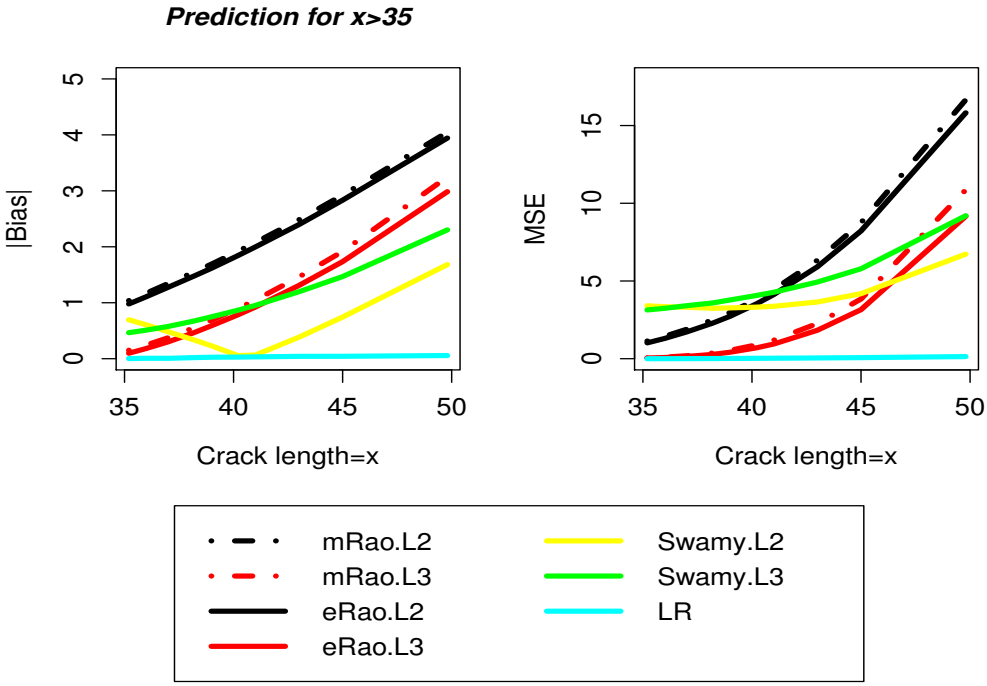

Figure 15: Prediction results of the linear models L2 and L3 using the first 131 observations of the new series

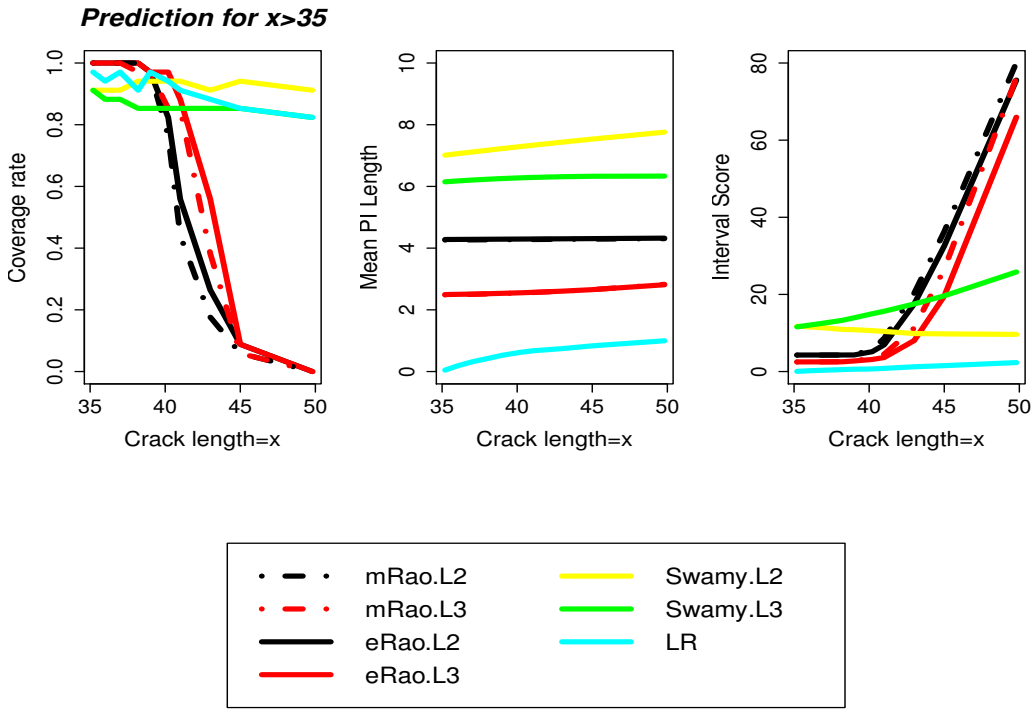

Figure 16: Prediction intervals of the linear models L2 and L3 using the first 131 observations of the new series

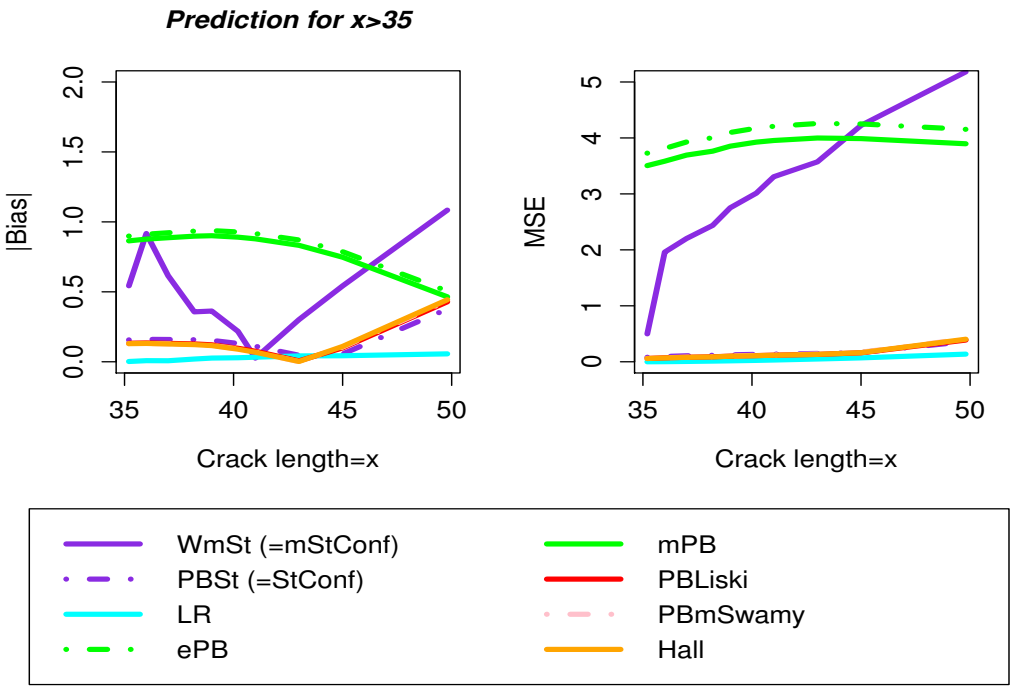

Figure 17: Prediction results of the nonlinear model using the first 131 observations of the new series

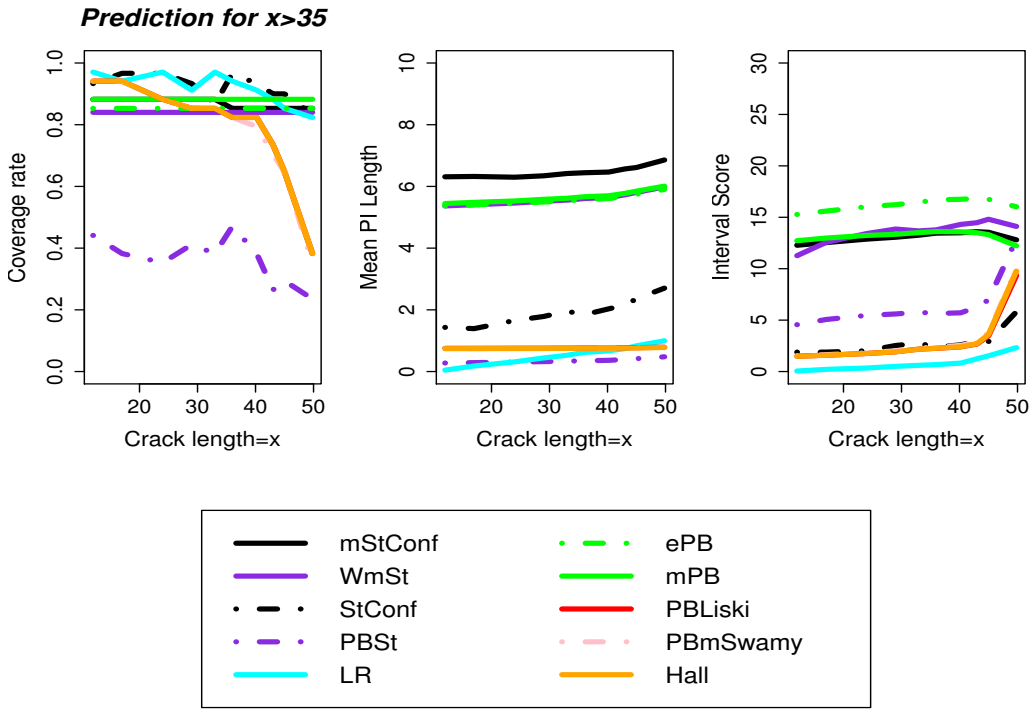

Figure 18: Prediction intervals of the nonlinear model using the first 131 observations of the new series

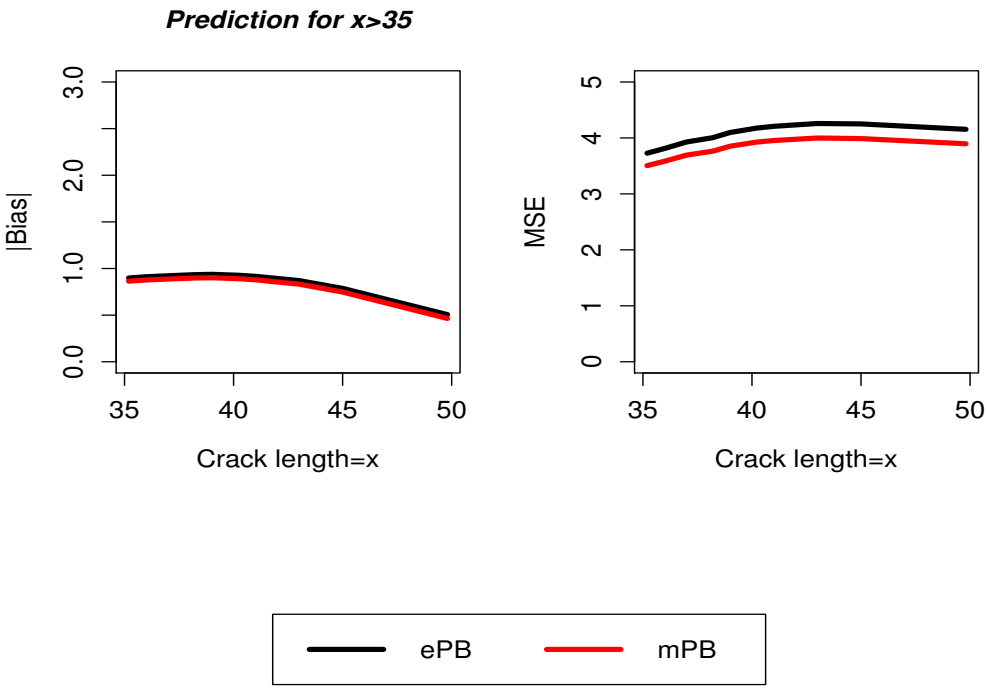

Figure 19: Prediction results of methods ePB and mPB using the first 131 observations of the new series

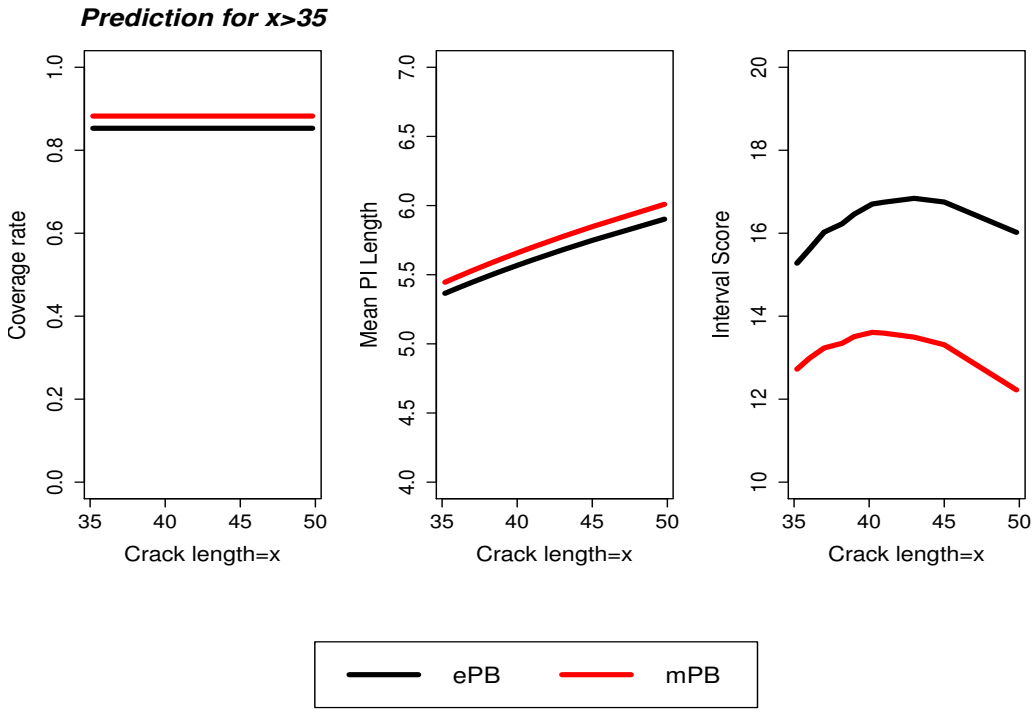

Figure 20: Prediction intervals of methods ePB and mPB using the first 131 observations of the new series

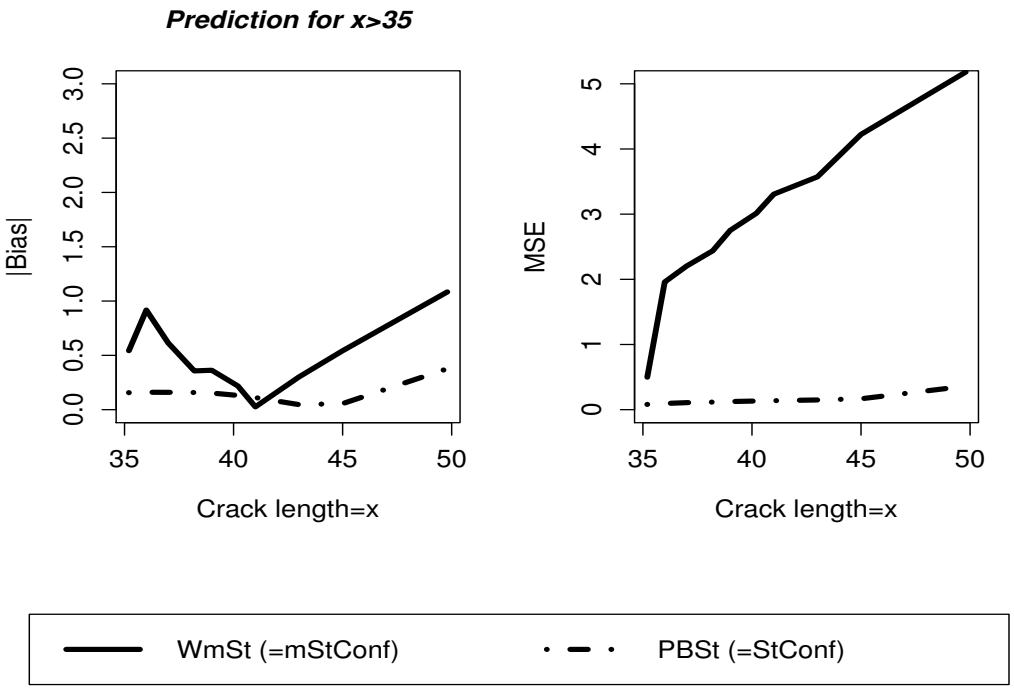

Figure 21: Prediction results of methods WmSt and PBSt using the first 131 observations of the new series

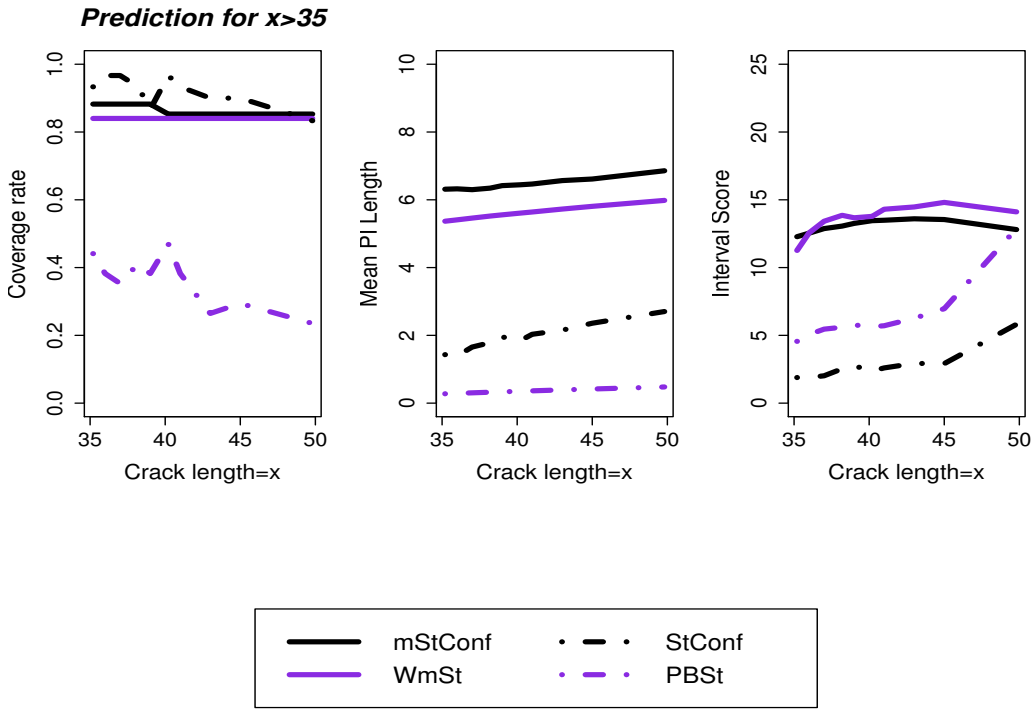

Figure 22: Prediction intervals of methods WmSt, PBSt, StConf and mStConf using the first 131 observations of the new series

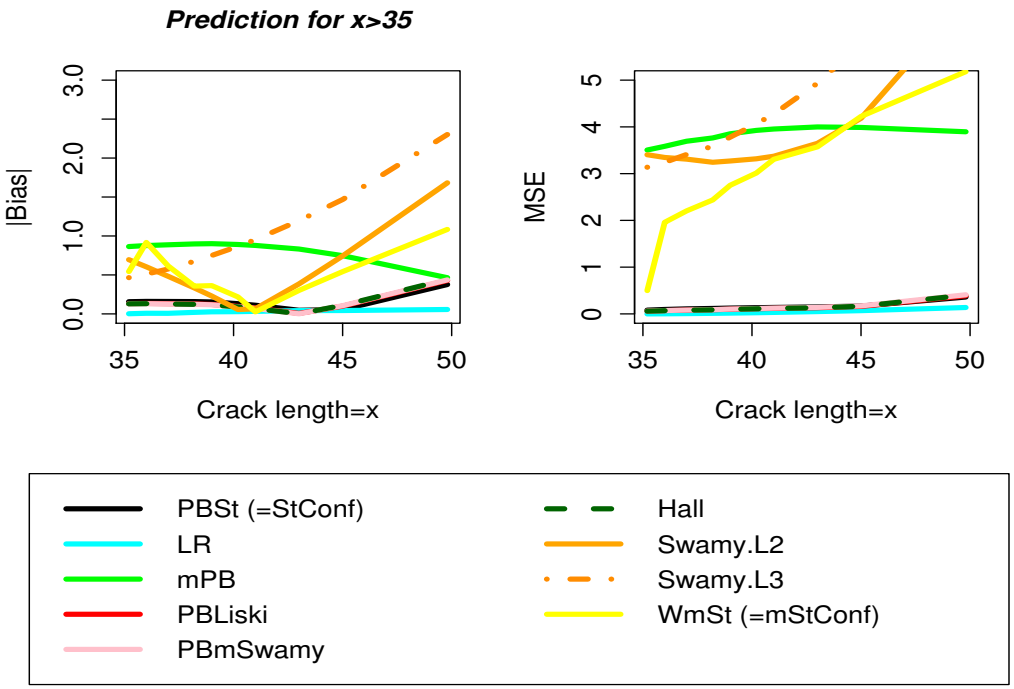

Figure 23: Overall comparison of prediction results for $x > 35$

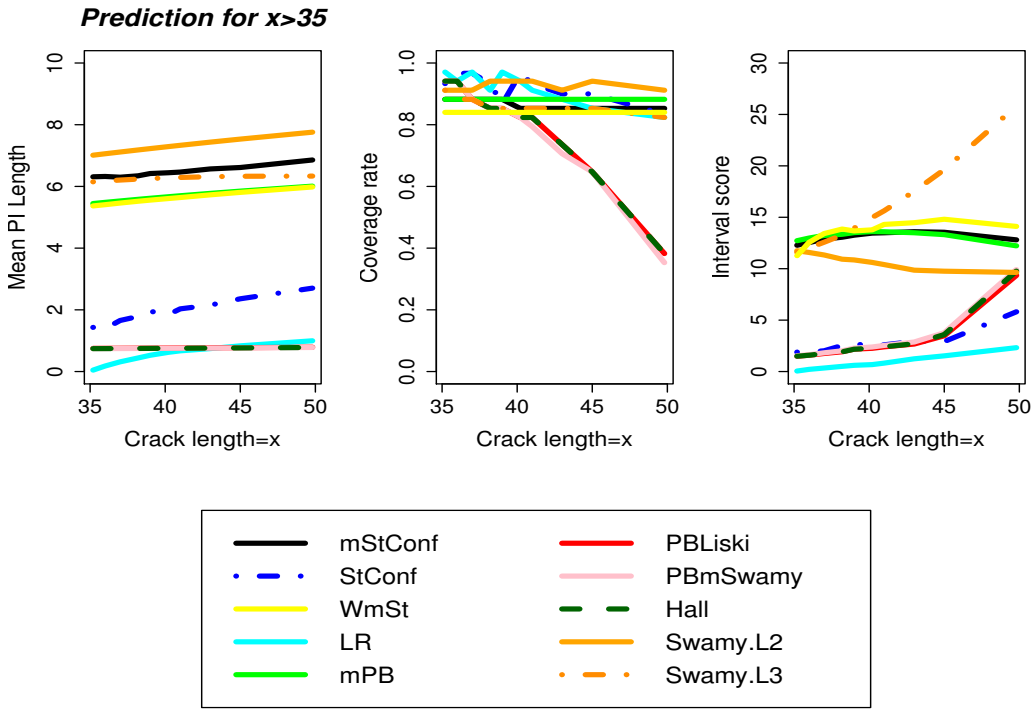

Figure 24: Overall comparison of prediction intervals for $x > 35$

6 Discussion and Conclusion

We proposed an exact method based on the MCEM algorithm and, an approximate method for estimating parameters in nonlinear random effects model. Based on the simulation results, among the discussed methods, method PBLiski gives closer estimations to the true ones for the variance covariance components of random effects. In terms of the fixed effects, the extension of Walker's estimation and Pinheiro and Bates' algorithm provide better estimations among the others (see Tables 2 and 3).

According to the obtained parameter estimations for the Virkler's data (see Table 8), it can be seen that the differences between the obtained estimates from the different methods are smaller than the ones in the simulation study. This implies that by increasing the number of observations within each individual, the proposed exact and approximate methods provide relatively the same estimation results. Moreover, from the simulation results, estimation procedures based on the near future prediction perform slightly better than the far away future prediction. In terms of computation time, it is clear that the approximate methods are usually faster than the exact method (the extension of Walker's estimation method). However, the computation time of our MCEM algorithm is readily outweighed by 3 advantages: (1) exact maximum likelihood estimations are achieved; (2) the estimates can be applied to assess the certainty of approximate maximum likelihood estimations; (3) one can determine that under what conditions and for which models the obtained estimates from the approximate methods are reliable.

According to the obtained prediction results from the simulation, methods WmSt, PBSt, PBLiski, Hall and LR give the better predictions among the others. Moreover, from the simulation results related to prediction interval, the best prediction intervals are obtained from methods PBLiski, Hall, LR and StConf. Generally, methods ePB and mPB provide relatively large prediction intervals with the high coverage rate and interval score. Moreover, method StConf works better than mStConf. In addition, for the far away future, method WmSt gives better prediction intervals than method PBSt and for the nearby future, method PBSt works better.

For the Virkler data, among the prediction methods for the linear models, Swamy's method generally presents better predictions and prediction intervals (the lower bias and MSE values for predictions and the smaller interval score for prediction intervals) than methods eRao and mRao. Bondeson (1990) states that the optimal weighted predictor is always better than simple predictors (e.g. Swamy's predictor) and proves that Rao's prediction is an optimal weighted predictor when Σ and σ^2 are known and mentions that it is impossible to prove the optimality of Rao's prediction when these parameters are replaced by their estimates, however,

it can be gussed that the predictor based on the estimations of parameter is also an optimal one. In such a case, Rao (1975) states that the prediction error will be larger than when parameters are known and will depend on how close the estimations to their true values. In our case, the extended and modified versions of Rao's prediction provided poor results. Using more accurate estimation methods can improve the results. When few observations from the new series are available, the estimate of parameters in method mRao is slightly different from the same one in method eRao such that it results in various prediction results. However, when more observations from the new series are available, it can be easily proved that the estimates of parameters in method mRao are approximately the same with the estimates from method eRao. And hence, there is no considerable difference between the results of methods eRao and mRao. It should be noted that the information from the new series is applied in the both methods, eRao and mRao, (see formulas of the prediction and prediction interval for both methods in Section 4.1.1) and the difference is only in the calculation of parameter estimates such that in the Rao's prediction only the information from the old series is used for the estimation of parameters whereas in the modified Rao's prediction, the information from the old and new series is used in the estimation of parameters. This difference can creat the different results.

In terms of prediction procedures for the nonlinear model, we obtained the following results. Generally, the prediction approach of Stirnemann et al. (which is used in methods PBSt and StConf) gives better predictions (the lower bias and MSE values) than the modified one (which is used in methods WmSt and mStConf). Stirnemann et al. (2011) prove that $\mathbb{E}(Y_0^F|y_0^P)$ is convergent to the conditional expectation of $g(A_0, x_0^F)$ given an infinite amount of observations of the new series. And then based on the simulation results, they show that this conditional expectation and the true unknown y_0^F are the same. Hence, when the number of available data in the new series increases, the estimation of $\mathbb{E}(Y_0^F|y_0^P)$ is a better predictor of the true unknown y_0^F. In methods PBSt (or StConf) and WmSt (or mStConf), we estimate $\mathbb{E}(Y_0^F|y_0^P)$ using the Monte Carlo approximation via MCMC algorithm and the ratio of two Monte Carlo approximations, respectively. And based on the obtained results, it seems that the Monte Carlo approximation via MCMC algorithm works better than the ratio of two Monte Carlo approximations in estimating $\mathbb{E}(Y_0^F|y_0^P)$. About the prediction intervals, generally methods WmSt and mStConf give the larger prediction intervals than methods PBSt and StConf. The reason is that, in methods WmSt and mStConf we simulate the random effect from its multidimensional normal distribution without any restriction. However, in methods PBSt and StConf we simulate the random effect from the conditional distribution of random effect given the past measurements from the new series. Restricting the random effect by the past measurements from the new

series can cause that the range of random sample changes to be not as large as the random sample from the nonconditional distribution of random effect. Therefore, in methods WmSt and mStConf the values of quantiles can be very large or very small and it can create large prediction intervals. Among the prediction intervals based on the confidence sets of parameters, generally method StConf works better than method mStConf. And between prediction intervals based on the parameter estimate, generally method WmSt performs better than method PBSt.
Moreover, the modified Pinheiro and Bates' prediction (mPB) provides better predictions and prediction intervals (the lower bias and MSE values for predictions and the smaller interval score for prediction intervals) than extension of the method of Pinheiro and Bates (ePB), especially when the number of available observations from the new series increases. This is because in the modified version, the estimates of parameters are obtained by using more information (the information from the old and new series) than the extended one. In the extended version of Pinheiro and Bates' prediction, only the information from the old series is used for the estimation of parameters and in fact the information from the new series is never used.
Among the linearization-based methods (i.e., PBLiski, PBmSwamy and Hall), the predictor in method PBLiski, which is an EM-predictor, provides better predictions. Because, based on the simulation, Liski and Nummi (1996) prove that their EM-predictor improves predictions in comparison to the non-iterative predictors. When few observations from the new seires is available, method PBLiski presents better prediction intervals (the smaller interval score) than method Hall, especially for the far away future. It should be mentioned that these methods give relatively similar prediction intervals and the main difference is in how the parameters are estimated. Such that in method PBLiski an EM algorithm is used, and in method Hall, an iterative estimation procedure of Pinheiro and Bates is applied. Moreover, based on the obtained estimation results from the simulation, this EM algorithm gives the closer estimates to the true values than the other methods, especially for the estimation of variance components. Therefore, it is expected that method PBLiski performs better than method Hall.
When few observations from the new seires are available, method PBmSwamy gives the worst predictions (with the highest bias and MSE values) but provides the largest prediction intervals with the highest coveragte rates, and the smallest interval score among the other linearization-based methods. The reason for such a poor prediction result is that, $\hat{Z}_0^{F(new)}$ in the prediction formula of method PBmSwamy cannot predict well the future observations, due to the little available information from the new series. Consequently, $\hat{Y}_0^{F(new)}$ and finally $\hat{Y}_0^F$ cannot result in good predictions (see Section 4.2.2).
When more observations from the new seires are available, all the three linearization-

based methods provide relatively the same results in prediction and also in prediction intervals. The reason is that, as it was mentioned in Section 5.2.2, when more observations from the new series are used, the estimation procedures in methods mPB (which is used in method Hall), PBLiski and PBmSwamy provide very close estimations. As such, the similar predictions can be obtained. As it was mentioned, methods PBLiski and Hall give the relatively similar prediction intervals and the main difference is in how parameters are estimated. Now since using more observations from the new series, relatively the same estimation results are obtained, both methods provide the same results. In terms of method PBmSwamy, it should be mentioned when more observations from the new series are used $\hat{Z}_0^{F(new)}$ can predict well the future observations (due to the more available information from the new series) and therefore $\hat{Y}_0^{F(new)}$ and finally $\hat{Y}_0^F$ can result in good predictions and now based on obtained relatively the same estimation results from these three mentioned methods, it is expected that this method provides predictions and prediction interval results as well as methods PBLiski and Hall. Since all of these three methods are based on the linearization using the Taylor expansion, and this expansion works better for the near future than for the far away future, the results for the near future predictions are better than for the far away future predictions. Finally for method LR, we concluded that this method provides the more flat curves and the best results among the other methods for the larger length values. For the smaller length values, this method does not work as well as for the larger length values. Since observations at the smaller length values contain less information of the old series than observations at the larger length values. Note that this method has the disadvantage that it can used only for the balanced data sets. It should be mentioned the presented models for fitting to Virkler's data are not the best because of the additive error term which can result in large prediction intervals even for the near future predictions.

In this research as in many other research works, the distribution of random effects and errors is considered to be normal. As it was mentioned by Meza et al. (2012), these assumptions can make inferences sensetive to the existence of outliers. Therefore, the use of heavy-tailed multivariate distributions like Student-t, the contaminated normal and slash can result in robust inferences. Several authors consider heavy-tailed distributions for random effects and errors for accomodating outliers (Welch and Richardson, 1997; Pinheiro et al., 2001; Yeap and Davidian, 2001; Lin and Lee, 2006; Staudenmayer et al., 2009). Thus, a generalization of this research would be to study prediction procedures for linear and nonlinear mixed models using heavy-tailed distributions and to present robust prediction intervals. This generalization can be an extension of Lin and Lee (2006) and Staudenmayer et al. (2009) estimation methods in linear mixed models and also an extension of

Yeap and Davidian (2001) and Meza et al. (2012) estimation methods in nonlinear mixed models. Another generalization of this research can be the extension of Mathew's method (Mathew et al., 2016) for calculating the prediction interval in a linear mixed model with one random effect to a linear mixed model with more than one random effects, and also to a nonlinear mixed model.

7 Appendix

7.1 Proof of theorems and lemmas

7.1.1 Proof of Theorem 2.3

Proof. The prediction error can be calculated as follows:

$$\hat{Y}_0^F - Y_0^F = \boldsymbol{x}_0^F \hat{\boldsymbol{\alpha}} - (\boldsymbol{x}_0^F \boldsymbol{\alpha} + E_0^F) = \boldsymbol{x}_0^F (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - E_0^F.$$

From (2.4) and the independency between $\hat{\boldsymbol{\alpha}}$ and E_0^F (because $\hat{\boldsymbol{\alpha}}$ is a linear function of Y_0^P and Y_0^P and E_0^F are independent) we have

$$\mathbb{E}(\hat{Y}_0^F - Y_0^F) = \boldsymbol{x}_0^F \mathbb{E}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - \mathbb{E}(E_0^F) = 0,$$

$$\mathbb{V}\mathrm{ar}(\hat{Y}_0^F - Y_0^F) = \boldsymbol{x}_0^F \mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\alpha}}) \boldsymbol{x}_0^{F\intercal} + \sigma^2$$
$$= \boldsymbol{x}_0^F \sigma^2 (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{x}_0^{F\intercal} + \sigma^2 = \sigma^2 [1 + \boldsymbol{x}_0^F (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{x}_0^{F\intercal}].$$

Moreover, since from (2.4) $\hat{\boldsymbol{\alpha}}$ has normal distribution and also E_0^F has also normal distribution then $\hat{Y}_0^F - Y_0^F$, which is a linear function of $\hat{\boldsymbol{\alpha}}$ and E_0^F, has also normal distribution.
For getting the prediction interval for Y_0^F, firstly we need to calculate the distribution of $(Y_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}})^\intercal (Y_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}})$. We define $P = \boldsymbol{X}_0^P (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{X}_0^{P\intercal}$ which is idempotent and symmetric and also we have $P\boldsymbol{X}_0^P = \boldsymbol{X}_0^P$. Then we have

$$\begin{aligned} &Y_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}} = \boldsymbol{X}_0^P \boldsymbol{\alpha} + E_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}} = -\boldsymbol{X}_0^P (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + E_0^P \\ &= -\boldsymbol{X}_0^P (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{X}_0^{P\intercal} E_0^P + E_0^P = (\boldsymbol{I}_{N_0} - \boldsymbol{X}_0^P (\boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P)^{-1} \boldsymbol{X}_0^{P\intercal}) E_0^P \\ &=: A E_0^P. \end{aligned} \tag{7.1}$$

From (7.1) and idempotent property of A, the residual sum of squares can be written as

$$(Y_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}})^\intercal (Y_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}}) = (A E_0^P)^\intercal (A E_0^P) = E_0^{P\intercal} A E_0^P = E_0^{P\intercal} (\boldsymbol{I}_{N_0} - P) E_0^P$$
$$= Y_0^{P\intercal} (\boldsymbol{I}_{N_0} - P) Y_0^P - Y_0^{P\intercal} \boldsymbol{X}_0^P \boldsymbol{\alpha} + Y_0^{P\intercal} P \boldsymbol{X}_0^P \boldsymbol{\alpha} - \boldsymbol{\alpha}^\intercal \boldsymbol{X}_0^{P\intercal} Y_0^P$$
$$+ \boldsymbol{\alpha}^\intercal \boldsymbol{X}_0^{P\intercal} P Y_0^P + \boldsymbol{\alpha}^\intercal \boldsymbol{X}_0^{P\intercal} \boldsymbol{X}_0^P \boldsymbol{\alpha} - \boldsymbol{\alpha}^\intercal \boldsymbol{X}_0^{P\intercal} P \boldsymbol{X}_0^P \boldsymbol{\alpha}.$$

Since $P\boldsymbol{X}_0^P = \boldsymbol{X}_0^P$, hence we have

$$(Y_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}})^\intercal (Y_0^P - \boldsymbol{X}_0^P \hat{\boldsymbol{\alpha}}) = Y_0^{P\intercal} (\boldsymbol{I}_{N_0} - P) Y_0^P = Y_0^{P\intercal} A Y_0^P.$$

Because A is idempotent and symmetric, then based on Theorem 2.2 we have

$$(Y_0^P - \boldsymbol{X}_0^P\hat{\boldsymbol{\alpha}})^\intercal(Y_0^P - \boldsymbol{X}_0^P\hat{\boldsymbol{\alpha}}) \sim \sigma^2\chi^2_{r,\lambda}. \tag{7.2}$$

r and λ can be calculated as

$$r = \text{rank}(A) = \text{tr}(A) = N_0 - q, \tag{7.3}$$

and

$$\sigma^2\lambda = \boldsymbol{\alpha}^\intercal\boldsymbol{X}_0^{P\intercal}A\boldsymbol{X}_0^P\boldsymbol{\alpha} = \boldsymbol{\alpha}^\intercal\boldsymbol{X}_0^{P\intercal}[\boldsymbol{X}_0^P\boldsymbol{\alpha} - \boldsymbol{X}_0^P(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P\boldsymbol{\alpha}] = 0$$

$$\Rightarrow \lambda = 0. \tag{7.4}$$

Finally, from (7.2), (7.3) and (7.4) it is concluded that

$$(Y_0^P - \boldsymbol{X}_0^P\hat{\boldsymbol{\alpha}})^\intercal(Y_0^P - \boldsymbol{X}_0^P\hat{\boldsymbol{\alpha}}) \sim \sigma^2\chi^2_{N_0-q}. \tag{7.5}$$

By use of (2.6) and (7.5) we can construct a t-statistic as follows:

$$\frac{\frac{\hat{Y}_0^F - Y_0^F}{\sqrt{\sigma^2[1+\boldsymbol{x}_0^F(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{x}_0^{F\intercal}]}}}{\sqrt{\frac{(Y_0^P-\boldsymbol{X}_0^P\hat{\boldsymbol{\alpha}})^\intercal(Y_0^P-\boldsymbol{X}_0^P\hat{\boldsymbol{\alpha}})}{\sigma^2(N_0-q)}}} = \frac{\frac{\hat{Y}_0^F - Y_0^F}{\sqrt{\sigma^2[1+\boldsymbol{x}_0^F(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{x}_0^{F\intercal}]}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}}$$

$$= \frac{\hat{Y}_0^F - Y_0^F}{\sqrt{\hat{\sigma}^2[1+\boldsymbol{x}_0^F(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{x}_0^{F\intercal}]}} \sim t_{N_0-q}.$$

Now, the $(1-\alpha)$-prediction interval for Y_0^F can be obtained by

$$PI = (\hat{y}_0^F \pm t_{N_0-q}(\alpha/2)\sqrt{\hat{\sigma}^2[1+\boldsymbol{x}_0^F(\boldsymbol{X}_0^{P\intercal}\boldsymbol{X}_0^P)^{-1}\boldsymbol{x}_0^{F\intercal}]}).$$

□

7.1.2 Proof of Theorem 2.4

Proof. By use of the first order Taylor expansion of $g(\hat{\boldsymbol{\alpha}}, x_0^F)$ around $\boldsymbol{\alpha}$, the prediction error $(\hat{Y}_0^F - Y_0^F)$ can approximately given by

$$\hat{Y}_0^F - Y_0^F = g(\hat{\boldsymbol{\alpha}}, x_0^F) - [g(\boldsymbol{\alpha}, x_0^F) + E_0^F] \approx g(\boldsymbol{\alpha}, x_0^F) + \tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - g(\boldsymbol{\alpha}, x_0^F) - E_0^F$$

$$= \tilde{\boldsymbol{x}}^F_{0,\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - E_0^F.$$

Because $\hat{\boldsymbol{\alpha}}$ is obtained based on Y_0^P and Y_0^P and E_0^F are independent therefore, $\hat{\boldsymbol{\alpha}}$ and E_0^F are also independent. Then from (2.13) we approximately have

$$\mathbb{E}(\hat{Y}_0^F - Y_0^F) \approx 0,$$

and

$$\begin{aligned}\mathbb{V}\text{ar}(\hat{Y}_0^F - Y_0^F) &\approx \sigma^2 \tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F (\tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P\intercal} \tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^P)^{-1} \tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^{F\intercal} + \sigma^2 \\ &= \sigma^2 [\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F (\tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P\intercal} \tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^P)^{-1} \tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^{F\intercal} + 1].\end{aligned}$$

Finally, it is concluded that

$$\hat{Y}_0^F - Y_0^F \approx \mathbb{N}(0, \sigma^2 [\tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^F (\tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^{P\intercal} \tilde{\boldsymbol{X}}_{0,\boldsymbol{\alpha}}^P)^{-1} \tilde{\boldsymbol{x}}_{0,\boldsymbol{\alpha}}^{F\intercal} + 1]).$$

By replacing $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$ in (2.15), the $(1-\alpha)$-prediction interval for Y_0^F is given by

$$PI = (\hat{y}_0^F \pm q_{\mathbb{N}(0,1),1-\alpha/2} \sqrt{\hat{\sigma}^2 [\tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^F (\tilde{\boldsymbol{X}}_{0,\hat{\boldsymbol{\alpha}}}^{P\intercal} \tilde{\boldsymbol{X}}_{0,\hat{\boldsymbol{\alpha}}}^P)^{-1} \tilde{\boldsymbol{x}}_{0,\hat{\boldsymbol{\alpha}}}^{F\intercal} + 1]}).$$

□

7.1.3 Proof of Lemma 3.1

Proof. Multiply (3.4) by $(\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \sigma^2\boldsymbol{I}_N)$, then we have

$$\begin{aligned}&(\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal+\sigma^2\boldsymbol{I}_N)[\sigma^{-2}\boldsymbol{I}_N-\sigma^{-2}\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal+\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1} \\ &\times(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal] \\ &= \sigma^{-2}\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal - \sigma^{-2}\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal + \boldsymbol{X}\Sigma\boldsymbol{X}^\intercal\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1} \\ &\times(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal+\boldsymbol{I}_N-\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal+\sigma^2\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \\ &= \sigma^{-2}\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal - \sigma^{-2}\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \boldsymbol{X}\Sigma(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \\ &+\boldsymbol{I}_N-\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal+\sigma^2\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \\ &= \boldsymbol{I}_N-\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal+(\boldsymbol{X}\Sigma+\sigma^2\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \\ &= \boldsymbol{I}_N-\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal+\boldsymbol{X}(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \\ &= \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal + \boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \\ &= \boldsymbol{I}_N.\end{aligned}$$

□

7.1.4 Proof of Theorem 3.1

Proof. Let $\boldsymbol{Y} := (Y_1^\intercal, \cdots, Y_I^\intercal)^\intercal$ and $\mathbb{X} := (\boldsymbol{X}^\intercal, \cdots, \boldsymbol{X}^\intercal)^\intercal$ be respectively the $(NI \times 1)$ vector of observations and the $(NI \times q)$ matrix of design matrices from the all individuals. Moreover, consider $D(\boldsymbol{X}) = \boldsymbol{I}_I \otimes \boldsymbol{X}$ the $(NI \times Iq)$ matrix, $\boldsymbol{\delta} := (\boldsymbol{\delta}_1^\intercal, \cdots, \boldsymbol{\delta}_I^\intercal)$ the $(Iq \times 1)$ vector of random effects and $\boldsymbol{E} := (E_1^\intercal, \cdots, E_I^\intercal)^\intercal$ the $(NI \times 1)$ vector of errors. With these notations, the RCR model for the all individuals is given by

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_I \end{pmatrix} = \begin{pmatrix} \boldsymbol{X} \\ \vdots \\ \boldsymbol{X} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_q \end{pmatrix} + \begin{pmatrix} \boldsymbol{X} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X} \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}_1 \\ \vdots \\ \boldsymbol{\delta}_I \end{pmatrix} + \begin{pmatrix} E_1 \\ \vdots \\ E_I \end{pmatrix}$$

$$\Leftrightarrow \boldsymbol{Y} = \mathbb{X}\boldsymbol{\alpha} + D(\boldsymbol{X})\boldsymbol{\delta} + \boldsymbol{E}. \tag{7.6}$$

From [A4], it is concluded that $\boldsymbol{\delta}$ and $\boldsymbol{E}$ are independent. In addition, from [A2], [A3] and [A5] we have $\boldsymbol{\delta} \sim \mathbb{N}_{Iq}(\boldsymbol{0}, \boldsymbol{I}_I \otimes \Sigma)$ and $\boldsymbol{E} \sim \mathbb{N}_{NI}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{NI})$. If we define $\boldsymbol{u} := D(\boldsymbol{X})\boldsymbol{\delta} + \boldsymbol{E}$ then

$$\boldsymbol{u} \sim \mathbb{N}_{NI}(\boldsymbol{0}, V),$$

$$\begin{aligned} V := \mathbb{C}\mathrm{ov}(D(\boldsymbol{X})\boldsymbol{\delta} + \boldsymbol{E}) &= D(\boldsymbol{X})\mathbb{C}\mathrm{ov}(\boldsymbol{\delta})D(\boldsymbol{X})^\intercal + \mathbb{C}\mathrm{ov}(\boldsymbol{E}) \\ &= (\boldsymbol{I}_I \otimes \boldsymbol{X})(\boldsymbol{I}_I \otimes \Sigma)(\boldsymbol{I}_I \otimes \boldsymbol{X})^\intercal + \sigma^2 \boldsymbol{I}_{NI} \\ &= \boldsymbol{I}_I \otimes (\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \sigma^2 \boldsymbol{I}_N). \end{aligned}$$

By use of $\boldsymbol{u}$, we can rewrite the RCR model (7.6) as a generalized linear model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\alpha} + \boldsymbol{u}.$$

In generalized linear models, $\hat{\boldsymbol{\alpha}}_{GLS} = (\mathbb{X}^\intercal V^{-1}\mathbb{X})^{-1}\mathbb{X}^\intercal V^{-1}\boldsymbol{Y}$ is the best linear unbiased estimation of $\boldsymbol{\alpha}$. Kadiyala and Oberhelman (1982) prove that $\hat{\boldsymbol{\alpha}}_{GLS} = \frac{1}{I}\sum_{i=1}^{I} \hat{A}_{i,OLS}$. By use of Lemma 3.1 we have

$$\begin{aligned} &\hat{\boldsymbol{\alpha}}_{GLS} = (\mathbb{X}^\intercal V^{-1}\mathbb{X})^{-1}\mathbb{X}^\intercal V^{-1}\boldsymbol{Y} \\ &= \big(\sum_{i=1}^{I} \boldsymbol{X}^\intercal(\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \sigma^2\boldsymbol{I}_N)^{-1}\boldsymbol{X}\big)^{-1} \sum_{i=1}^{I} \boldsymbol{X}^\intercal(\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \sigma^2\boldsymbol{I}_N)^{-1} Y_i \\ &= \big(I\boldsymbol{X}^\intercal(\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \sigma^2\boldsymbol{I}_N)^{-1}\boldsymbol{X}\big)^{-1} \boldsymbol{X}^\intercal(\boldsymbol{X}\Sigma\boldsymbol{X}^\intercal + \sigma^2\boldsymbol{I}_N)^{-1} \sum_{i=1}^{I} Y_i \\ &= \frac{1}{I}[\boldsymbol{X}^\intercal\big(\sigma^{-2}\boldsymbol{I}_N - \sigma^{-2}\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal + \boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}(\Sigma + \sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1} \end{aligned}$$

$$
\begin{aligned}
&\times(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\big)\boldsymbol{X}]^{-1}\boldsymbol{X}^\intercal\big(\sigma^{-2}\boldsymbol{I}_N-\sigma^{-2}\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\\
&+\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\big)\sum_{i=1}^{I}Y_i\\
&=\frac{1}{I}[\sigma^{-2}\boldsymbol{X}^\intercal\boldsymbol{X}-\sigma^{-2}\boldsymbol{X}^\intercal\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{X}+\boldsymbol{X}^\intercal\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}\\
&\times(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{X}]^{-1}\big(\sigma^{-2}\boldsymbol{X}^\intercal-\sigma^{-2}\boldsymbol{X}^\intercal\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\\
&+\boldsymbol{X}^\intercal\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\big)\sum_{i=1}^{I}Y_i\\
&=\frac{1}{I}[(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}]^{-1}(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\sum_{i=1}^{I}Y_i\\
&=\frac{1}{I}(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})(\Sigma+\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\sum_{i=1}^{I}Y_i\\
&=\frac{1}{I}\sum_{i=1}^{I}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal Y_i\\
&=\frac{1}{I}\sum_{i=1}^{I}\hat{A}_{i,OLS}.
\end{aligned}
\tag{7.7}
$$

□

7.1.5 Proof of Theorem 3.2

Proof. From the classic linear regression model theory, $\hat{\sigma}_i^2$ for $i \in \{1, \cdots, I\}$ is the unbiased and consistent estimation of σ^2. Hence $\hat{\sigma}^2$ is also an unbiased and consistent estimation of σ^2, since

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}(\frac{1}{I}\sum_{i=1}^{I}\hat{\sigma}_i^2) = \frac{1}{I}\sum_{i=1}^{I}\mathbb{E}(\hat{\sigma}_i^2) = \sigma^2,$$

and

$$\text{plim}_{N\to\infty}(\hat{\sigma}^2) = \frac{1}{I}\sum_{i=1}^{I}\text{plim}_{N\to\infty}(\hat{\sigma}_i^2) = \sigma^2.$$

The expectation of $\hat{\Sigma}$ is given by

$$\mathbb{E}\big(\frac{1}{I-1}\sum_{i=1}^{I}(\hat{A}_{i,OLS}-\hat{\boldsymbol{\alpha}}_{GLS})(\hat{A}_{i,OLS}-\hat{\boldsymbol{\alpha}}_{GLS})^{\intercal}-\hat{\sigma}^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\big)$$

$$=\frac{1}{I-1}\sum_{i=1}^{I}\mathbb{E}\big((\hat{A}_{i,OLS}-\boldsymbol{\alpha}+\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}_{GLS})(\hat{A}_{i,OLS}-\boldsymbol{\alpha}+\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}_{GLS})^{\intercal}\big)$$

$$-\mathbb{E}(\hat{\sigma}^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1})$$

$$=\frac{1}{I-1}\sum_{i=1}^{I}\mathbb{E}\big((\hat{A}_{i,OLS}-\boldsymbol{\alpha})(\hat{A}_{i,OLS}-\boldsymbol{\alpha})^{\intercal}+(\hat{A}_{i,OLS}-\boldsymbol{\alpha})(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}_{GLS})^{\intercal}$$

$$+(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}_{GLS})(\hat{A}_{i,OLS}-\boldsymbol{\alpha})^{\intercal}+(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}_{GLS})(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}_{GLS})^{\intercal}\big)-\mathbb{E}(\hat{\sigma}^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1})$$

$$=\frac{1}{I-1}[I\mathbb{E}((\hat{A}_{i,OLS}-\boldsymbol{\alpha})(\hat{A}_{i,OLS}-\boldsymbol{\alpha})^{\intercal})-I\mathbb{E}((\hat{\boldsymbol{\alpha}}_{GLS}-\boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}}_{GLS}-\boldsymbol{\alpha})^{\intercal})$$

$$-I\mathbb{E}((\hat{\boldsymbol{\alpha}}_{GLS}-\boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}}_{GLS}-\boldsymbol{\alpha})^{\intercal})+I\mathbb{E}((\hat{\boldsymbol{\alpha}}_{GLS}-\boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}}_{GLS}-\boldsymbol{\alpha})^{\intercal})]-\mathbb{E}(\hat{\sigma}^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1})$$

$$=\frac{1}{I-1}[I\mathbb{C}\mathrm{ov}(\hat{A}_{i,OLS})-I\mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\alpha}}_{GLS})]-\sigma^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}$$

$$=\frac{1}{I-1}[I(\Sigma+\sigma^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1})-I\frac{1}{I}(\Sigma+\sigma^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1})]-\sigma^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}$$

$$=\Sigma+\sigma^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}-\sigma^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}=\Sigma.$$

Therefore, $\hat{\Sigma}$ is an unbiased estimation of Σ. For studying the consistency of $\hat{\Sigma}$, we suppose that $B:=\lim_{N\to\infty}(N(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1})$ exists and is positive definite hence we have

$$\mathrm{plim}_{N\to\infty}(\hat{\sigma}^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1})=\mathrm{plim}_{N\to\infty}(\hat{\sigma}^2)\lim_{N\to\infty}(\frac{1}{N})\lim_{N\to\infty}(N(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1})$$

$$=\sigma^2\cdot 0\cdot B=0. \tag{7.8}$$

Swamy (1971) proves that the matrix $\hat{S}:=\sum_{i=1}^{I}(\hat{A}_{i,OLS}-\hat{\alpha}_{GLS})(\hat{A}_{i,OLS}-\hat{\alpha}_{GLS})^{\intercal}$ for $N\to\infty$ converges in probability to the matrix $S:=\sum_{i=1}^{I}(A_i-\bar{A})(A_i-\bar{A})^{\intercal}$, where $\bar{A}$ is the mean of A_i's. Now by use of this convergency we have

$$\mathrm{plim}_{N\to\infty}\hat{\Sigma}=\mathrm{plim}_{N\to\infty}(\frac{\hat{S}}{I-1}-\hat{\sigma}^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1})$$

$$=\mathrm{plim}_{N\to\infty}(\frac{\hat{S}}{I-1})-\mathrm{plim}_{N\to\infty}(\hat{\sigma}^2(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1})$$

$$= \frac{S}{I-1},$$

because the empirical covariance matrix is always a consistent estimation of covariance matrix, hence if $I \to \infty$ then

$$\text{plim}_{I\to\infty} \frac{S}{I-1} = \Sigma,$$

and finally,

$$\hat{\Sigma} \xrightarrow{N\to\infty} \frac{S}{I-1} \xrightarrow{I\to\infty} \Sigma. \tag{7.9}$$

□

7.1.6 Proof of Lemma 3.2

Proof. Because $\hat{Y}_0^F$ and Y_0^F have normal distribution, therefore $(\hat{Y}_0^F - Y_0^F)$ has also normal distribution. For the mean and variance of prediction error we have

$$\begin{aligned}
\mathbb{E}(\hat{Y}_0^F - Y_0^F) &= \mathbb{E}(\hat{Y}_0^F) - \mathbb{E}(Y_0^F) = \mathbb{E}(\boldsymbol{x}_0^F \hat{\boldsymbol{\alpha}}_{GLS}) - \mathbb{E}(\boldsymbol{x}_0^F A_0 + E_0^F) \\
&= \boldsymbol{x}_0^F \boldsymbol{\alpha} - \boldsymbol{x}_0^F \boldsymbol{\alpha} = 0. \quad \text{(from (3.5) and [A2],[A3])} \\
\mathbb{V}\text{ar}(\hat{Y}_0^F - Y_0^F) &= \mathbb{V}\text{ar}(\hat{Y}_0^F) + \mathbb{V}\text{ar}(Y_0^F) \quad \text{(since } Y_0^F \text{ independent of } \boldsymbol{Y}) \\
&= \mathbb{V}\text{ar}(\boldsymbol{x}_0^F \hat{\boldsymbol{\alpha}}_{GLS}) + \mathbb{V}\text{ar}(\boldsymbol{x}_0^F A_0 + E_0^F) \\
&= \boldsymbol{x}_0^F \mathbb{C}\text{ov}(\hat{\boldsymbol{\alpha}}_{GLS})(\boldsymbol{x}_0^F)^\intercal + \boldsymbol{x}_0^F \mathbb{C}\text{ov}(A_0)(\boldsymbol{x}_0^F)^\intercal + \mathbb{V}\text{ar}(E_0^F) \quad \text{(from [A4])} \\
&= \frac{1}{I} \boldsymbol{x}_0^F (\Sigma + \sigma^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal + \boldsymbol{x}_0^F \Sigma (\boldsymbol{x}_0^F)^\intercal + \sigma^2. \quad \text{(from (3.5) and [A2], [A3])}
\end{aligned}$$

□

7.1.7 Proof of Theorem 3.3

Proof. Since $\hat{\Sigma}$ and $\hat{\sigma^2}$ are unbiased, therefore $\hat{\nu}$ is also an unbiased estimation of ν. Now, the asymptotic properties of $\hat{\nu}$ should be studied. Assume $B := \lim_{N\to\infty}(N(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1})$ exists and is positive definite, then if $N, I \to \infty$, we have

$$\hat{\nu}^* := \text{plim}_{N\to\infty}(\hat{\nu}) = \text{plim}_{N\to\infty}(\frac{1}{I} \boldsymbol{x}_0^F (\hat{\Sigma} + \hat{\sigma}^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal + \boldsymbol{x}_0^F \hat{\Sigma} (\boldsymbol{x}_0^F)^\intercal + \hat{\sigma}^2)$$

$$\begin{aligned}
&= \frac{1}{I} \boldsymbol{x}_0^F \text{plim}_{N\to\infty}(\hat{\Sigma})(\boldsymbol{x}_0^F)^\intercal + \frac{1}{I} \boldsymbol{x}_0^F \text{plim}_{N\to\infty}(\hat{\sigma}^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal \\
&\quad + \boldsymbol{x}_0^F \text{plim}_{N\to\infty}(\hat{\Sigma})(\boldsymbol{x}_0^F)^\intercal + \text{plim}_{N\to\infty}(\hat{\sigma}^2)
\end{aligned}$$

$$= \frac{1}{I}\boldsymbol{x}_0^F(\frac{1}{I-1}S)(\boldsymbol{x}_0^F)^\intercal + \boldsymbol{x}_0^F(\frac{1}{I-1}S)(\boldsymbol{x}_0^F)^\intercal + \sigma^2 (\text{from (7.9) and the consistency of } \hat{\sigma}^2)$$

$$= \frac{I+1}{I}\boldsymbol{x}_0^F(\frac{1}{I-1}S)(\boldsymbol{x}_0^F)^\intercal + \sigma^2$$

$$\begin{aligned}\text{plim}_{I\to\infty}(\hat{\nu}^*) &= \text{plim}_{I\to\infty}(\frac{I+1}{I}\boldsymbol{x}_0^F(\frac{1}{I-1}S)(\boldsymbol{x}_0^F)^\intercal + \sigma^2) \\ &= \lim_{I\to\infty}(\frac{I+1}{I})\boldsymbol{x}_0^F \,\text{plim}_{I\to\infty}(\frac{1}{I-1}S)(\boldsymbol{x}_0^F)^\intercal + \sigma^2 \\ &= \boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + \sigma^2 \quad (\text{from (7.9)}).\end{aligned}$$

Consequently, $\hat{\nu}$ is a consistent estimation for $\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + \sigma^2$ but inconsistent for ν. For getting the asymptotic distribution of $\hat{Y}_0^F - Y_0^F$, study of the asymptotic properties of ν is required. If $N, I \to \infty$ then we have

$$\nu^* := \lim_{N\to\infty}\nu = \lim_{N\to\infty}(\frac{1}{I}\boldsymbol{x}_0^F(\Sigma + \sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal + \boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + \sigma^2)$$

$$= \frac{I+1}{I}\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + (\frac{\sigma^2}{I})\boldsymbol{x}_0^F\lim_{N\to\infty}(\frac{1}{N})\lim_{N\to\infty}(N(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})(\boldsymbol{x}_0^F)^\intercal + \sigma^2$$

$$= \frac{I+1}{I}\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + (\frac{\sigma^2}{I})\boldsymbol{x}_0^F(0\cdot B)(\boldsymbol{x}_0^F)^\intercal + \sigma^2$$

$$= \frac{I+1}{I}\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + \sigma^2$$

$$\begin{aligned}\lim_{I\to\infty}\nu^* &= \lim_{I\to\infty}(\frac{I+1}{I})\boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + \sigma^2 \\ &= \boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + \sigma^2.\end{aligned}$$

Finally by applying the Slutzky theorem, the asymptotic distribution of prediction error is given by

$$\frac{\hat{Y}_0^F - Y_0^F}{\sqrt{\hat{\nu}}} \xrightarrow[N,I\to\infty]{D} \frac{Z}{\sqrt{\mathbb{V}\text{ar}(Z)}} \sim \mathbb{N}(0,1),$$

where $Z \sim \mathbb{N}(0, \boldsymbol{x}_0^F\Sigma(\boldsymbol{x}_0^F)^\intercal + \sigma^2)$. In conclusion, the approximate $(1-\alpha)$-prediction interval for Y_0^F from a new individual is given by

$$[\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}].$$

□

7.1.8 Proof of Lemma 4.3

Proof. For getting the prediction interval, we need to obtain the distribution of prediction error. The prediction error is given by

$$\tilde{Y}_0^F - Y_0^F = \boldsymbol{x}_0^F \tilde{\boldsymbol{\alpha}} + V_0^{FP} V_0^{P^{-1}} (Y_0^P - \boldsymbol{X}_0^P \tilde{\boldsymbol{\alpha}}) - \boldsymbol{x}_0^F A_0 + E_0^F.$$

Since the prediction error is a linear combination of the three normal variables ($\tilde{\boldsymbol{\alpha}}$, A_0 and E_0^F), hence it has also normal distribution. For the expectation of prediction error from (4.25) and the assumptions for random errors and random effects we have

$$\mathbb{E}(\tilde{Y}_0^F - Y_0^F) = \mathbb{E}(\boldsymbol{x}_0^F \tilde{\boldsymbol{\alpha}} + V_0^{FP} V_0^{P^{-1}} (Y_0^P - \boldsymbol{X}_0^P \tilde{\boldsymbol{\alpha}}) - \boldsymbol{x}_0^F A_0 + E_0^F)$$

$$= \boldsymbol{x}_0^F \mathbb{E}(\tilde{\boldsymbol{\alpha}}) + V_0^{FP} V_0^{P^{-1}} (\mathbb{E}(Y_0^P) - \boldsymbol{X}_0^P \mathbb{E}(\tilde{\boldsymbol{\alpha}})) - \boldsymbol{x}_0^F \mathbb{E}(A_0) + \mathbb{E}(E_0^F)$$

$$= \boldsymbol{x}_0^F \boldsymbol{\alpha} + V_0^{FP} V_0^{P^{-1}} (\boldsymbol{X}_0^P \boldsymbol{\alpha} - \boldsymbol{X}_0^P \boldsymbol{\alpha}) - \boldsymbol{x}_0^F \boldsymbol{\alpha} + 0$$

$$= 0.$$

By use of (4.24), the prediction error can be simplified as

$$\tilde{Y}_0^F - Y_0^F = V_0^{FP} (V_0^P)^{-1} (Y_0^P - \boldsymbol{X}_0^P \tilde{\boldsymbol{\alpha}}) - (Y_0^F - \boldsymbol{x}_0^F \tilde{\boldsymbol{\alpha}})$$

$$= V_0^{FP} (V_0^P)^{-1} (Y_0^P - \boldsymbol{X}_0^P \tilde{\boldsymbol{\alpha}} + \boldsymbol{X}_0^P \boldsymbol{\alpha} - \boldsymbol{X}_0^P \boldsymbol{\alpha}) - (Y_0^F - \boldsymbol{x}_0^F \tilde{\boldsymbol{\alpha}} + \boldsymbol{x}_0^F \boldsymbol{\alpha} - \boldsymbol{x}_0^F \boldsymbol{\alpha})$$

$$= V_0^{FP} (V_0^P)^{-1} [(Y_0^P - \boldsymbol{X}_0^P \boldsymbol{\alpha}) + (\boldsymbol{X}_0^P \boldsymbol{\alpha} - \boldsymbol{X}_0^P \tilde{\boldsymbol{\alpha}})] - [(Y_0^F - \boldsymbol{x}_0^F \boldsymbol{\alpha}) + (\boldsymbol{x}_0^F \boldsymbol{\alpha} - \boldsymbol{x}_0^F \tilde{\boldsymbol{\alpha}})].$$

Define matrices A, A_1, A_2 and B, B_1, B_2 as

$$A := A_1 + A_2$$

where

$$A_1 := V_0^{FP} (V_0^P)^{-1} [Y_0^P - \boldsymbol{X}_0^P \boldsymbol{\alpha}], \quad A_2 := V_0^{FP} (V_0^P)^{-1} [\boldsymbol{X}_0^P \boldsymbol{\alpha} - \boldsymbol{X}_0^P \tilde{\boldsymbol{\alpha}}],$$

and

$$B := B_1 + B_2$$

where

$$B_1 := Y_0^F - \boldsymbol{x}_0^F \boldsymbol{\alpha}, \quad B_2 := \boldsymbol{x}_0^F \boldsymbol{\alpha} - \boldsymbol{x}_0^F \tilde{\boldsymbol{\alpha}}.$$

Now we have

$$\mathbb{V}\mathrm{ar}[\tilde{Y}_0^F - Y_0^F] = \mathbb{V}\mathrm{ar}[A - B] = \mathbb{V}\mathrm{ar}(A) + \mathbb{V}\mathrm{ar}(B) - \mathbb{C}\mathrm{ov}(A, B) - \mathbb{C}\mathrm{ov}(B, A).$$

Also

$$\mathbb{V}\mathrm{ar}(A) = \mathbb{V}\mathrm{ar}(A_1)+\mathbb{V}\mathrm{ar}(A_2)+\mathbb{C}\mathrm{ov}(A_1, A_2)+(\mathbb{C}\mathrm{ov}(A_1, A_2))^\intercal.$$

Since (by use of the independency between Y_0^P and $Y_i, i = 1, \cdots, I$),

$$\begin{aligned}\mathbb{C}\mathrm{ov}(Y_0^P, \tilde{\boldsymbol{\alpha}}) &= \mathbb{C}\mathrm{ov}(Y_0^P, (\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P)^{-1}\boldsymbol{X}_0^{P\intercal}V_0^{P^{-1}}Y_0^P)\\ &= V_0^P[V_0^P]^{-1}\boldsymbol{X}_0^P(\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P)^{-1} = \boldsymbol{X}_0^P(\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P)^{-1},\end{aligned}$$

where $\boldsymbol{X}_P := (\boldsymbol{X}_1^\intercal, \cdots, \boldsymbol{X}_I^\intercal, \boldsymbol{X}_0^{P\intercal})^\intercal$, $V_P := \mathbb{V}\mathrm{ar}(\boldsymbol{Y}_{obs}) = \boldsymbol{I}_{I+1} \otimes (\boldsymbol{X}_i^P\Sigma\boldsymbol{X}_i^{P\intercal} + \sigma^2\boldsymbol{I}_{N_i})$ with $\boldsymbol{Y}_{obs} := (Y_1^\intercal, \cdots, Y_I^\intercal, Y_0^{P\intercal})^\intercal$ and $\boldsymbol{X}_i^P := \boldsymbol{X}_i, \quad i \in \{1, \cdots, I\}$, hence we have

$$\begin{aligned}&\mathrm{Var}(A) = V_0^{FP}(V_0^P)^{-1}V_0^P(V_0^P)^{-1}(V_0^{FP})^\intercal\\ &+V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}(V_0^{FP})^\intercal\\ &-V_0^{FP}(V_0^P)^{-1}\mathbb{C}\mathrm{ov}(Y_0^P, \tilde{\boldsymbol{\alpha}})(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}(V_0^{FP})^\intercal\\ &-(V_0^{FP}(V_0^P)^{-1}\mathbb{C}\mathrm{ov}(Y_0^P, \tilde{\boldsymbol{\alpha}})(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}(V_0^{FP})^\intercal)^\intercal\\ &= V_0^{FP}(V_0^P)^{-1}(V_0^{FP})^\intercal+V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}(V_0^{FP})^\intercal\\ &-V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P(\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P)^{-1}(\boldsymbol{X}_0^P)^\intercal[V_0^P]^{-1}[V_0^{FP}]^\intercal\\ &-V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P(\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P)^{-1}(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}[V_0^{FP}]^\intercal.\end{aligned} \tag{7.10}$$

Moreover,

$$\begin{aligned}\mathbb{V}\mathrm{ar}(B) &= \mathbb{V}\mathrm{ar}(B_1)+\mathbb{V}\mathrm{ar}(B_2)+\mathbb{C}\mathrm{ov}(B_1, B_2)+\mathbb{C}\mathrm{ov}(B_2, B_1)\\ &= \mathbb{V}\mathrm{ar}(B_1) + \mathbb{V}\mathrm{ar}(B_2) + \mathbb{C}\mathrm{ov}(B_1, B_2) + (\mathbb{C}\mathrm{ov}(B_1, B_2))^\intercal.\end{aligned}$$

Because $\mathbb{C}\mathrm{ov}(Y_0^F, \tilde{\boldsymbol{\alpha}}) = V_0^{FP}[V_0^P]^{-1}\boldsymbol{X}_0^P[\boldsymbol{X}_P^\intercal(V_P)^{-1}\boldsymbol{X}_P]^{-1}$, thus we have

$$\begin{aligned}&\mathbb{V}\mathrm{ar}(B) = V_0^F+\boldsymbol{x}_0^F\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})(\boldsymbol{x}_0^F)^\intercal-V_0^{FP}[V_0^P]^{-1}\boldsymbol{X}_0^P[\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P]^{-1}[\boldsymbol{x}_0^F]^\intercal\\ &\quad-\boldsymbol{x}_0^F[\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P]^{-1}[\boldsymbol{X}_0^P]^\intercal(V_0^P)^{-1}[V_0^{FP}]^\intercal,\end{aligned} \tag{7.11}$$

In addition

$$\begin{aligned}&-\mathbb{C}\mathrm{ov}(A, B) = -\mathbb{C}\mathrm{ov}(A_1+A_2, B_1+B_2)\\ &= -[\mathbb{C}\mathrm{ov}(A_1, B_1)+\mathbb{C}\mathrm{ov}(A_1, B_2)+\mathbb{C}\mathrm{ov}(A_2, B_1)+\mathbb{C}\mathrm{ov}(A_2, B_2)]\\ &= -\mathbb{C}\mathrm{ov}(V_0^{FP}(V_0^P)^{-1}Y_0^P, Y_0^F)+\mathbb{C}\mathrm{ov}(V_0^{FP}(V_0^P)^{-1}Y_0^P, \boldsymbol{x}_0^F\tilde{\boldsymbol{\alpha}})\\ &+\mathbb{C}\mathrm{ov}(V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P\tilde{\boldsymbol{\alpha}}, Y_0^F)-\mathbb{C}\mathrm{ov}(V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P\tilde{\boldsymbol{\alpha}}, \boldsymbol{x}_0^F\tilde{\boldsymbol{\alpha}})\end{aligned}$$

$$
\begin{aligned}
&= -V_0^{FP}(V_0^P)^{-1}(V_0^{FP})^\intercal + V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P[\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P]^{-1}(\boldsymbol{x}_0^F)^\intercal \\
&+ V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P[\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P]^{-1}(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}(V_0^{FP})^\intercal \\
&- V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})(\boldsymbol{x}_0^F)^\intercal.
\end{aligned} \tag{7.12}
$$

By the fact $(\mathbb{C}\mathrm{ov}(A,B))^\intercal = \mathbb{C}\mathrm{ov}(B,A)$, we have

$$
\begin{aligned}
-\mathbb{C}\mathrm{ov}(B,A) = &-V_0^{FP}(V_0^P)^{-1}(V_0^{FP})^\intercal + \boldsymbol{x}_0^F[\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P]^{-1}(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}(V_0^{FP})^\intercal \\
&+ V_0^{FP}[V_0^P]^{-1}\boldsymbol{X}_0^P[\boldsymbol{X}_P^\intercal V_P^{-1}\boldsymbol{X}_P]^{-1}(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}(V_0^{FP})^\intercal \\
&- \boldsymbol{x}_0^F\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}(V_0^{FP})^\intercal.
\end{aligned} \tag{7.13}
$$

Finally by replacing (7.10)-(7.13) in the variance of prediction error and removing the same terms with different signs, it is concluded that

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}[\tilde{Y}_0^F - Y_0^F] = &\, V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}(V_0^{FP})^\intercal + V_0^F + \boldsymbol{x}_0^F\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})(\boldsymbol{x}_0^F)^\intercal \\
&- V_0^{FP}(V_0^P)^{-1}(V_0^{FP})^\intercal - V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})(\boldsymbol{x}_0^F)^\intercal \\
&- \boldsymbol{x}_0^F\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})(\boldsymbol{X}_0^P)^\intercal(V_0^P)^{-1}(V_0^{FP})^\intercal.
\end{aligned}
$$

The obtained result can also be written as follows:

$$
\mathbb{V}\mathrm{ar}[\tilde{Y}_0^F - Y_0^F] = \Omega_0^F + M_0[\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\alpha}})]M_0^\intercal =: \Omega_0,
$$

where

$$
\Omega_0^F = V_0^F - V_0^{FP}(V_0^P)^{-1}(V_0^{FP})^\intercal, \quad M_0 = \boldsymbol{x}_0^F - V_0^{FP}(V_0^P)^{-1}\boldsymbol{X}_0^P.
$$

Hence, the $(1-\alpha)$-prediction interval for Y_0^F can be given by

$$
\tilde{y}_0^F \pm q_{1-\alpha/2}\sqrt{\Omega_0}.
$$

□

7.1.9 Proof of Theorem 4.3

Proof. Based on (4.50), because $\hat{\boldsymbol{\alpha}}$, $\boldsymbol{\delta}_0$ and $E_0 := \begin{pmatrix} E_0^P \\ E_0^F \end{pmatrix}$ are distributed as normal, hence $Y_0^F - \hat{Y}_0^F$ is also approximately distributed as normal. From (4.46) and (4.50), the mean of prediction error is easily calculated as

$$
\mathbb{E}(Y_0^F - \hat{Y}_0^F) \approx (-\widehat{V_0}^{FP}\widehat{V_0}^{P^{-1}}, 1)[\tilde{\boldsymbol{X}}_0\mathbb{E}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{X}}_0\mathbb{E}(\boldsymbol{\delta}_0) + \mathbb{E}(E_0)] \approx 0.
$$

The estimated variance of prediction error is given by

$$\widehat{\mathbb{V}\mathrm{ar}}(Y_0^F - \hat{Y}_0^F) \approx (-\widehat{V_0}^{FP}\widehat{V_0}^{P^{-1}}, 1)\widehat{\mathbb{C}\mathrm{ov}}[\tilde{\boldsymbol{X}}_0(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{X}}_0\boldsymbol{\delta}_0 + E_0]\begin{pmatrix}(-\widehat{V_0}^{FP}\widehat{V_0}^{P^{-1}})^\intercal \\ 1\end{pmatrix}$$

$$=: \hat{\nu}. \tag{7.14}$$

From the independency between $\boldsymbol{\delta}_0$ and E_0, the $\widehat{\mathbb{C}\mathrm{ov}}[\tilde{\boldsymbol{X}}_0(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{X}}_0\boldsymbol{\delta}_0 + E_0]$ is given by

$$\widehat{\mathbb{C}\mathrm{ov}}[\tilde{\boldsymbol{X}}_0(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{X}}_0\boldsymbol{\delta}_0 + E_0] = \tilde{\boldsymbol{X}}_0\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}})\tilde{\boldsymbol{X}}_0^\intercal + \tilde{\boldsymbol{X}}_0\hat{\Sigma}\tilde{\boldsymbol{X}}_0^\intercal + \hat{\sigma}^2\boldsymbol{I}_{N_0+1}$$

$$-\tilde{\boldsymbol{X}}_0\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\delta}_0)\tilde{\boldsymbol{X}}_0^\intercal - \tilde{\boldsymbol{X}}_0\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}}, E_0) - \tilde{\boldsymbol{X}}_0\widehat{\mathbb{C}\mathrm{ov}}(\boldsymbol{\delta}_0, \hat{\boldsymbol{\alpha}})\tilde{\boldsymbol{X}}_0^\intercal - \widehat{\mathbb{C}\mathrm{ov}}(E_0, \hat{\boldsymbol{\alpha}})\tilde{\boldsymbol{X}}_0^\intercal. \tag{7.15}$$

From (4.46), (4.44) and the independency between Z_i and $\boldsymbol{\delta}_0$ and E_0 we can calculate the following covariances as:

$$\begin{aligned}
\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\delta}_0) &= [\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^\intercal\widehat{V_i}^{-1}\tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^P]^{-1}\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\widehat{\mathbb{C}\mathrm{ov}}(Z_0^P, \boldsymbol{\delta}_0) \\
&\approx [\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^\intercal\widehat{V_i}^{-1}\tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^P]^{-1}\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\widehat{\mathbb{C}\mathrm{ov}}(\tilde{\boldsymbol{X}}_0^P(\boldsymbol{\alpha}+\boldsymbol{\delta}_0) + E_0^P, \boldsymbol{\delta}_0) \\
&= [\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^\intercal\widehat{V_i}^{-1}\tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^P]^{-1}\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^P\hat{\Sigma}.
\end{aligned} \tag{7.16}$$

$$\begin{aligned}
\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}}, E_0) &= [\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^\intercal\widehat{V_i}^{-1}\tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^P]^{-1}\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\widehat{\mathbb{C}\mathrm{ov}}(Z_0^P, \begin{pmatrix}E_0^P \\ E_0^F\end{pmatrix}) \\
&\approx [\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^\intercal\widehat{V_i}^{-1}\tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^P]^{-1}\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\widehat{\mathbb{C}\mathrm{ov}}(\tilde{\boldsymbol{X}}_0^P(\boldsymbol{\alpha}+\boldsymbol{\delta}_0) + E_0^P, \begin{pmatrix}E_0^P \\ E_0^F\end{pmatrix}) \\
&= [\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^\intercal\widehat{V_i}^{-1}\tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^P]^{-1}\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\widehat{\mathbb{C}\mathrm{ov}}(E_0^P, \begin{pmatrix}E_0^P \\ E_0^F\end{pmatrix}) \\
&= [\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^\intercal\widehat{V_i}^{-1}\tilde{\boldsymbol{X}}_i + \tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^P]^{-1}\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}(\hat{\sigma}^2\boldsymbol{I}_{N_0}, \boldsymbol{0}_{N_0\times 1}).
\end{aligned} \tag{7.17}$$

By replacing (7.16) and (7.17) in (7.15) we have

$$\widehat{\mathbb{C}\mathrm{ov}}[\tilde{\boldsymbol{X}}_0(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}) + \tilde{\boldsymbol{X}}_0\boldsymbol{\delta}_0 + E_0] \approx \tilde{\boldsymbol{X}}_0\widehat{\mathbb{C}\mathrm{ov}}(\hat{\boldsymbol{\alpha}})\tilde{\boldsymbol{X}}_0^\intercal + \tilde{\boldsymbol{X}}_0\hat{\Sigma}\tilde{\boldsymbol{X}}_0^\intercal + \hat{\sigma}^2\boldsymbol{I}_{N_0+1}$$

$$-\tilde{\boldsymbol{X}}_0[\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^{\intercal}\widehat{V}_i^{-1}\tilde{\boldsymbol{X}}_i+\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}]^{-1}[\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}\hat{\Sigma}\tilde{\boldsymbol{X}}_0^{\intercal}+\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}$$

$$\times(\hat{\sigma}^2\boldsymbol{I}_{N_0},\boldsymbol{0}_{N_0\times 1})]-[\tilde{\boldsymbol{X}}_0\hat{\Sigma}\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}+(\hat{\sigma}^2\boldsymbol{I}_{N_0},\boldsymbol{0}_{N_0\times 1})^{\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}]$$

$$\times[\sum_{i=1}^{I}\tilde{\boldsymbol{X}}_i^{\intercal}\widehat{V}_i^{-1}\tilde{\boldsymbol{X}}_i+\tilde{\boldsymbol{X}}_0^{P\intercal}\widehat{V_0}^{P^{-1}}\tilde{\boldsymbol{X}}_0^{P}]^{-1}\tilde{\boldsymbol{X}}_0^{\intercal}.$$

Now, the estimated variance of prediction error in (7.14) can be calculated. The $(1-\alpha)$-prediction interval for Y_0^F is given by

$$[\hat{y}_0^F - q_{1-\alpha/2}\sqrt{\hat{\nu}}, \hat{y}_0^F + q_{1-\alpha/2}\sqrt{\hat{\nu}}].$$

□

7.1.10 Proof of Lemma 4.4

Proof.

$$\int f(y_0^F|a_0,\hat{\boldsymbol{\beta}})f(a_0|y_0^P,\hat{\boldsymbol{\beta}})da_0 = \int \frac{f(y_0^F,a_0|\hat{\boldsymbol{\beta}})}{f(a_0|\hat{\boldsymbol{\beta}})}\frac{f(y_0^P,a_0|\hat{\boldsymbol{\beta}})}{f(y_0^P|\hat{\boldsymbol{\beta}})}da_0$$

$$= \frac{1}{f(y_0^P|\hat{\boldsymbol{\beta}})}\int \frac{f(y_0^F|a_0,\hat{\boldsymbol{\beta}})f(a_0|\hat{\boldsymbol{\beta}})}{f(a_0|\hat{\boldsymbol{\beta}})}f(y_0^P|a_0,\hat{\boldsymbol{\beta}})f(a_0|\hat{\boldsymbol{\beta}})da_0$$

$$= \frac{1}{f(y_0^P|\hat{\boldsymbol{\beta}})}\int f(y_0^F|a_0,\hat{\boldsymbol{\beta}})f(y_0^P|a_0,\hat{\boldsymbol{\beta}})f(a_0|\hat{\boldsymbol{\beta}})da_0$$

$$= \frac{1}{f(y_0^P|\hat{\boldsymbol{\beta}})}\int f(y_0^P,y_0^F|a_0,\hat{\boldsymbol{\beta}})f(a_0|\hat{\boldsymbol{\beta}})da_0$$

$$= \frac{1}{f(y_0^P|\hat{\boldsymbol{\beta}})}\int \frac{f(y_0^P,y_0^F,a_0|\hat{\boldsymbol{\beta}})}{f(a_0|\hat{\boldsymbol{\beta}})}f(a_0|\hat{\boldsymbol{\beta}})da_0$$

$$= \int \frac{f(y_0^P,y_0^F,a_0|\hat{\boldsymbol{\beta}})}{f(y_0^P|\hat{\boldsymbol{\beta}})}da_0$$

$$= \int f(y_0^F,a_0|y_0^P,\hat{\boldsymbol{\beta}})da_0$$

$$= f(y_0^F|y_0^P,\hat{\boldsymbol{\beta}}).$$

□

7.1.11 Proof of Lemma 4.5

Proof. By use of Lemma 4.4, we have

$$\mathbb{E}(Y_0^F|y_0^P,\hat{\boldsymbol{\beta}}) = \int y_0^F f(y_0^F|y_0^P,\hat{\boldsymbol{\beta}})dy_0^F = \int y_0^F \int f(y_0^F|a_0,\hat{\boldsymbol{\beta}})f(a_0|y_0^P,\hat{\boldsymbol{\beta}})da_0 dy_0^F$$

$$= \int\int y_0^F f(y_0^F|a_0,\hat{\boldsymbol{\beta}})dy_0^F f(a_0|y_0^P,\hat{\boldsymbol{\beta}})da_0.$$

The last expression can be simplified because $\int y_0^F f(y_0^F|a_0,\hat{\boldsymbol{\beta}})dy_0^F = \mathbb{E}(Y_0^F|a_0,\hat{\boldsymbol{\beta}}) = g(a_0,x_0^F)$. Therefore,

$$\mathbb{E}(Y_0^F|y_0^P,\hat{\boldsymbol{\beta}}) = \int g(a_0,x_0^F)f(a_0|y_0^P,\hat{\boldsymbol{\beta}})da_0.$$

Stirnemann et al. (2011) propose using the MCMC algorithm for simulating an M-sample $(\tilde{a}_0^{(1)},\cdots,\tilde{a}_0^{(M)})$ from $f(a_0|y_0^P,\hat{\boldsymbol{\beta}})$, by use of this random sample, the estimation of $\mathbb{E}(Y_0^F|y_0^P,\hat{\boldsymbol{\beta}})$ is given by

$$\hat{y}_0^F := \widehat{\mathbb{E}(Y_0^F|y_0^P,\hat{\boldsymbol{\beta}})} \approx \frac{1}{M}\sum_{m=1}^{M} g(\tilde{a}_0^{(m)},x_0^F).$$

The conditional expectation of $\mathbb{E}(A_0|y_0^P,\hat{\boldsymbol{\beta}})$ is given by

$$\mathbb{E}(A_0|y_0^P,\hat{\boldsymbol{\beta}}) = \int a_0 f(a_0|y_0^P,\hat{\boldsymbol{\beta}})da_0.$$

Therefore, the Monte Carlo estimation of $\mathbb{E}(A_0|y_0^P,\hat{\boldsymbol{\beta}})$ is given by

$$\hat{a}_0 := \widehat{\mathbb{E}(A_0|y_0^P,\hat{\boldsymbol{\beta}})} \approx \frac{1}{M}\sum_{m=1}^{M} \tilde{a}_0^{(m)}$$

where $(\tilde{a}_0^{(1)},\cdots,\tilde{a}_0^{(M)})$ is a random sample from $f(a_0|y_0^P,\hat{\boldsymbol{\beta}})$ using the MCMC algorithm. □

7.1.12 Proof of Theorem 4.4

Proof.

$$P_{\boldsymbol{\beta}}(Y_0^F \in \mathbb{P}(\boldsymbol{y}_{obs})) = P_{\boldsymbol{\beta}}(Y_0^F \in \cup_{\boldsymbol{\beta}\in\hat{\Theta}_\alpha(\boldsymbol{y})}[q_{\alpha/2}(\boldsymbol{\beta},y_0^P),q_{1-\alpha/2}(\boldsymbol{\beta},y_0^P)])$$

$$\geq P_{\boldsymbol{\beta}}(Y_0^F \in [q_{\alpha/2}(\boldsymbol{\beta},y_0^P),q_{1-\alpha/2}(\boldsymbol{\beta},y_0^P)] \wedge \quad \boldsymbol{\beta}\in\hat{\Theta}_\alpha(\boldsymbol{y})) \quad ,$$

now from the independency between Y_0^F and $\boldsymbol{Y}$ and also between Y_0^P and $\boldsymbol{Y}$ we have

$$P_{\boldsymbol{\beta}}(Y_0^F \in \mathbb{P}(\boldsymbol{y}_{obs})) \geq P_{\boldsymbol{\beta}}(Y_0^F \in [q_{\alpha/2}(\boldsymbol{\beta}, y_0^P), q_{1-\alpha/2}(\boldsymbol{\beta}, y_0^P)]) P_{\boldsymbol{\beta}}(\boldsymbol{\beta} \in \hat{\Theta}_\alpha(\boldsymbol{y}))$$

$\geq (1-\alpha)(1-\alpha) = (1-\alpha)^2$ from the definition of $(1-\alpha)$-confidence set (Müller et al., 2016). □

7.1.13 Proof of Theorem 4.6

Proof. At first, it should be noted that for any $\boldsymbol{\beta}_* \in \Theta$ we have

$$P_{\boldsymbol{\beta}_*}(Y_0^F \in [F^{-1}_{\boldsymbol{\beta}_*}(\alpha/2), F^{-1}_{\boldsymbol{\beta}_*}(1-\alpha/2)]) = F_{\boldsymbol{\beta}_*}(F^{-1}_{\boldsymbol{\beta}_*}(1-\alpha/2)) - F_{\boldsymbol{\beta}_*}(F^{-1}_{\boldsymbol{\beta}_*}(\alpha/2))$$

$$= 1 - \alpha/2 - \alpha/2 = 1 - \alpha.$$

Then for any $\boldsymbol{\beta}_* \in \Theta$ we obtain

$$P_{\boldsymbol{\beta}_*}(Y_0^F \notin \mathbb{P}(\boldsymbol{y}_{obs})) = P_{\boldsymbol{\beta}_*}(Y_0^F \notin \cup_{\boldsymbol{\beta} \in \hat{\Theta}(\boldsymbol{y}_{obs})} [F^{-1}_{\boldsymbol{\beta}}(\alpha/2), F^{-1}_{\boldsymbol{\beta}}(1-\alpha/2)])$$

$$= P_{\boldsymbol{\beta}_*}(Y_0^F \notin \cup_{\boldsymbol{\beta} \in \hat{\Theta}(\boldsymbol{y}_{obs})} [F^{-1}_{\boldsymbol{\beta}}(\alpha/2), F^{-1}_{\boldsymbol{\beta}}(1-\alpha/2)], \boldsymbol{\beta}_* \in \hat{\Theta}(\boldsymbol{y}_{obs}))$$

$$+ P_{\boldsymbol{\beta}_*}(Y_0^F \notin \cup_{\boldsymbol{\beta} \in \hat{\Theta}(\boldsymbol{y}_{obs})} [F^{-1}_{\boldsymbol{\beta}}(\alpha/2), F^{-1}_{\boldsymbol{\beta}}(1-\alpha/2)], \boldsymbol{\beta}_* \notin \hat{\Theta}(\boldsymbol{y}_{obs}))$$

$$\leq P_{\boldsymbol{\beta}_*}(Y_0^F \notin [F^{-1}_{\boldsymbol{\beta}_*}(\alpha/2), F^{-1}_{\boldsymbol{\beta}_*}(1-\alpha/2)]) + P_{\boldsymbol{\beta}_*}(\boldsymbol{\beta}_* \notin \hat{\Theta}(\boldsymbol{y}_{obs})) \leq 2\alpha$$

(Müller et al., 2016). □

7.2 Computational Methods of Pinheiro and Bates' estimation method

The proposed alternating algorithm of Lindstrom and Bates (1990) involves the LME step and PNLS step. In the LME step, the optimization of a linear mixed effects log-likelihood is needed. Two optimization procedures can be applied, the EM algorithm and the Newton-Raphson algorithm.

Estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}_i$:
At the PNLS step of the Lindstrom and Bates's alternating algorithm, the optimization of the penalized sum of squares (3.16) with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}_i$ is required. Pinheiro and Bates (2000) define the augmented response and the model function vectors at the $(k+1)$th iteration as follows:

$$\tilde{Y}_i := \begin{pmatrix} Y_i \\ \mathbf{0} \end{pmatrix}, \quad \tilde{g}_i(\boldsymbol{\delta}_i, \boldsymbol{\alpha}, \boldsymbol{x}_i) := \begin{pmatrix} g_i(\boldsymbol{\delta}_i + \boldsymbol{\alpha}, \boldsymbol{x}_i) \\ \Delta(\boldsymbol{\theta}^{(k)})\boldsymbol{\delta}_i \end{pmatrix}.$$

Then the penalized sum of squares (3.16) at the $(k+1)$th iteration can be represented as

$$\sum_{i=1}^{I} ||\tilde{Y}_i - \tilde{g}_i(\boldsymbol{\delta}_i, \boldsymbol{\alpha}, \boldsymbol{x}_i)||^2. \tag{7.18}$$

Based on (7.18), conditional on $\Delta(\boldsymbol{\theta}^{(k)})$, we can study the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}_i$ as a standard nonlinear least squares problem. For standard nonlinear least squares problems, the Gaussian-Newton method is commonly applied. In this method, at the $(k+1)$th iteration $\tilde{g}_i(\boldsymbol{\delta}_i, \boldsymbol{\alpha}, \boldsymbol{x}_i)$ is substituted by a first order Taylor expansion around the current estimates $\boldsymbol{\delta}_i^{(k)}$ and $\boldsymbol{\alpha}^{(k)}$ as

$$\tilde{g}_i(\boldsymbol{\delta}_i, \boldsymbol{\alpha}, \boldsymbol{x}_i) \approx \tilde{g}_i(\boldsymbol{\delta}_i^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{x}_i) + \hat{\tilde{\boldsymbol{W}}}_i^{(k+1)}(\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^{(k)}) + \hat{\tilde{\boldsymbol{X}}}_i^{(k+1)}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(k)}),$$

where

$$\hat{\tilde{\boldsymbol{X}}}_i^{(k+1)} := \frac{\partial \tilde{g}_i(\boldsymbol{\delta}_i, \boldsymbol{\alpha}, \boldsymbol{x}_i)}{\partial \boldsymbol{\alpha}^\intercal}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(k)}, \boldsymbol{\delta}_i=\boldsymbol{\delta}_i^{(k)}} = \begin{pmatrix} \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)} \\ \boldsymbol{0} \end{pmatrix},$$

$$\hat{\tilde{\boldsymbol{W}}}_i^{(k+1)} := \frac{\partial \tilde{g}_i(\boldsymbol{\delta}_i, \boldsymbol{\alpha}, \boldsymbol{x}_i)}{\partial \boldsymbol{\delta}_i^\intercal}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(k)}, \boldsymbol{\delta}_i=\boldsymbol{\delta}_i^{(k)}} = \begin{pmatrix} \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)} \\ \Delta(\boldsymbol{\theta}^{(k)}) \end{pmatrix},$$

with $\tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)} = \frac{\partial g_i(\boldsymbol{a}, \boldsymbol{x}_i)}{\partial \boldsymbol{a}^\intercal}\Big|_{\boldsymbol{a}=\boldsymbol{\alpha}^{(k)}+\boldsymbol{\delta}_i^{(k)}} \in \mathbb{R}^{N_i \times q}$. At each Gauss-Newton iteration, the following least squares problem should be solved with respect to $\boldsymbol{\delta}_i$ and $\boldsymbol{\alpha}$:

$$\sum_{i=1}^{I} ||[\tilde{Y}_i - \tilde{g}_i(\boldsymbol{\delta}_i^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{x}_i)] - \hat{\tilde{\boldsymbol{W}}}_i^{(k+1)}(\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^{(k)}) - \hat{\tilde{\boldsymbol{X}}}_i^{(k+1)}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(k)})||^2$$

$$= \sum_{i=1}^{I} || \begin{pmatrix} Y_i \\ \boldsymbol{0} \end{pmatrix} - \begin{pmatrix} g_i(\boldsymbol{\delta}_i^{(k)} + \boldsymbol{\alpha}^{(k)}, \boldsymbol{x}_i) \\ \Delta(\boldsymbol{\theta}^{(k)})\boldsymbol{\delta}_i^{(k)} \end{pmatrix} - \begin{pmatrix} \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)} \\ \Delta(\boldsymbol{\theta}^{(k)}) \end{pmatrix} (\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^{(k)})$$

$$- \begin{pmatrix} \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)} \\ \boldsymbol{0} \end{pmatrix} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(k)})||^2$$

$$= \sum_{i=1}^{I} || \begin{pmatrix} Y_i - g_i(\boldsymbol{\delta}_i^{(k)} + \boldsymbol{\alpha}^{(k)}, \boldsymbol{x}_i) - \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)}(\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^{(k)}) - \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(k)}) \\ \boldsymbol{0} - \Delta(\boldsymbol{\theta}^{(k)})\boldsymbol{\delta}_i^{(k)} - \Delta(\boldsymbol{\theta}^{(k)})(\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^{(k)}) - \boldsymbol{0} \end{pmatrix} ||^2$$

$$= \sum_{i=1}^{I} || \begin{pmatrix} Y_i - g_i(\boldsymbol{\delta}_i^{(k)} + \boldsymbol{\alpha}^{(k)}, \boldsymbol{x}_i) + \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)}(\boldsymbol{\delta}_i^{(k)} + \boldsymbol{\alpha}^{(k)}) - \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)}(\boldsymbol{\alpha} + \boldsymbol{\delta}_i) \\ -\Delta(\boldsymbol{\theta}^{(k)})\boldsymbol{\delta}_i \end{pmatrix} ||^2$$

$$= \sum_{i=1}^{I} || \begin{pmatrix} Z_i^{(k+1)} - \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)}(\boldsymbol{\alpha} + \boldsymbol{\delta}_i) \\ -\Delta(\boldsymbol{\theta}^{(k)})\boldsymbol{\delta}_i \end{pmatrix} ||^2$$

$$= \sum_{i=1}^{I} || \begin{pmatrix} Z_i^{(k+1)} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)} \\ \Delta(\boldsymbol{\theta}^{(k)}) \end{pmatrix} \boldsymbol{\delta}_i - \begin{pmatrix} \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)} \\ \mathbf{0} \end{pmatrix} \boldsymbol{\alpha} ||^2$$

$$\sum_{i=1}^{I} ||\tilde{\tilde{Z}}_i^{(k+1)} - \hat{\tilde{\boldsymbol{W}}}_i^{(k+1)} \boldsymbol{\delta}_i - \hat{\tilde{\boldsymbol{X}}}_i^{(k+1)} \boldsymbol{\alpha}||^2, \quad \text{where} \quad \tilde{\tilde{Z}}_i^{(k+1)} := \begin{pmatrix} Z_i^{(k+1)} \\ \mathbf{0} \end{pmatrix}, \tag{7.19}$$

with

$$Z_i^{(k+1)} := Y_i - g_i(\boldsymbol{\delta}_i^{(k)} + \boldsymbol{\alpha}^{(k)}, \boldsymbol{x}_i) + \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)}(\boldsymbol{\delta}_i^{(k)} + \boldsymbol{\alpha}^{(k)}) \approx \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)}(\boldsymbol{\delta}_i + \boldsymbol{\alpha}) + E_i.$$

Pinheiro and Bates (2000) propose use of the Ortohogonal triangular decompositions, which are called the QR decompositions, of rectangular matrices for solving least squares problems (Thisted, 1988). If H is an $n \times p$ matrix ($n \geq p$) of rank p, then the QR decomposition of H is

$$H = Q \begin{pmatrix} R \\ \mathbf{0} \end{pmatrix},$$

where Q is an $n \times n$- and orthogonal-matrix (i.e. $Q^\intercal Q = QQ^\intercal = \boldsymbol{I}$) and R is an $p \times p$- and upper triangular-matrix. The orthogonal matrix Q maintains norms of vectors under multiplication by Q or by $Q^\intercal$. Consider the QR decomposition of $\hat{\tilde{\boldsymbol{W}}}_i^{(k+1)}$ as

$$\hat{\tilde{\boldsymbol{W}}}_i^{(k+1)} = Q_{(i)} \begin{pmatrix} R_{11(i)} \\ \mathbf{0} \end{pmatrix}, \tag{7.20}$$

where $Q_{(i)}$ is an $(N_i + q) \times (N_i + q)$ orthogonal matrix and $R_{11(i)}$ is an $q \times q$ upper triangular matrix. Then because $Q_{(i)}$ is an orthogonal matrix we have

$$\sum_{i=1}^{I} ||\tilde{\tilde{Z}}_i^{(k+1)} - \hat{\tilde{\boldsymbol{W}}}_i^{(k+1)} \boldsymbol{\delta}_i - \hat{\tilde{\boldsymbol{X}}}_i^{(k+1)} \boldsymbol{\alpha}||^2 = \sum_{i=1}^{I} ||Q_{(i)}^\intercal (\tilde{\tilde{Z}}_i^{(k+1)} - \hat{\tilde{\boldsymbol{W}}}_i^{(k+1)} \boldsymbol{\delta}_i - \hat{\tilde{\boldsymbol{X}}}_i^{(k+1)} \boldsymbol{\alpha})||^2$$

$$= \sum_{i=1}^{I} ||c_{1(i)} - R_{10(i)} \boldsymbol{\alpha} - R_{11(i)} \boldsymbol{\delta}_i||^2 + ||c_{0(i)} - R_{00(i)} \boldsymbol{\alpha}||^2, \tag{7.21}$$

where the $q \times q$ matrix $R_{10(i)}$, the $N_i \times q$ matrix $R_{00(i)}$, the q-vector $c_{1(i)}$ and the N_i-vector $c_{0(i)}$ are determined by

$$\begin{pmatrix} R_{10(i)} \\ R_{00(i)} \end{pmatrix} = Q_{(i)}^\intercal \hat{\tilde{\boldsymbol{X}}}_i^{(k+1)} \text{and} \quad \begin{pmatrix} c_{1(i)} \\ c_{0(i)} \end{pmatrix} = Q_{(i)}^\intercal \tilde{\tilde{Z}}_i^{(k+1)}. \tag{7.22}$$

According to (7.20) and (7.22) at the $(k+1)$th iteration, the QR decomposition of an augmented matrix is given by

$$\begin{pmatrix} \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)} & \tilde{\tilde{\boldsymbol{X}}}_i^{(k+1)} & Z_i^{(k+1)} \\ \Delta(\boldsymbol{\theta}^{(k)}) & \mathbf{0} & \mathbf{0} \end{pmatrix} = Q_{(i)} \begin{pmatrix} R_{11(i)} & R_{10(i)} & c_{1(i)} \\ \mathbf{0} & R_{00(i)} & c_{0(i)} \end{pmatrix}. \tag{7.23}$$

Pinheiro and Bates (2000) consider another the QR decomposition

$$\begin{pmatrix} R_{00(1)} & c_{0(1)} \\ \vdots & \vdots \\ R_{00(I)} & c_{0(I)} \end{pmatrix} = Q_0 \begin{pmatrix} R_{00} & c_0 \\ \mathbf{0} & c_{-1} \end{pmatrix}. \tag{7.24}$$

By the fact that Q_0 is an orthogonal matrix, if the second term of (7.21) is multiplied by $Q_0^\intercal$, the equation (7.21) can be rewritten as

$$(\sum_{i=1}^{I} ||c_{1(i)} - R_{10(i)}\boldsymbol{\alpha} - R_{11(i)}\boldsymbol{\delta}_i||^2) + ||c_0 - R_{00}\boldsymbol{\alpha}||^2 + ||c_{-1}||^2. \tag{7.25}$$

By supposing that the R_{00} is full rank, the unique least squares estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}_i$ are given at the $(k+1)$th iteration by

$$\hat{\boldsymbol{\alpha}}^{(k+1)} := R_{00}^{-1} c_0, \tag{7.26}$$

$$\hat{\boldsymbol{\delta}}_i^{(k+1)} := R_{11(i)}^{-1}(c_{1(i)} - R_{10(i)}\hat{\boldsymbol{\alpha}}^{(k+1)}), \quad i = 1, \quad , I. \tag{7.27}$$

Pinheiro and Bates (2000) suggest to apply the step-halving to guarantee that the new estimates causes a decrease of the objective function (7.18). At the $(k+1)$th Gauss-Newton iteration for estimation of $\boldsymbol{\alpha}$ as an example, consider $d^{(k+1)} = \hat{\boldsymbol{\alpha}}^{(k+1)} - \boldsymbol{\alpha}^{(k)}$ as parameter increment. If the value of objective function related to $\boldsymbol{\alpha}^{(k+1)}$ is less than the value at $\boldsymbol{\alpha}^{(k)}$, then the value is preserved and the algorithm continues to the next step. Otherwise, the updated estimate at the $(k+1)$th iteration is set to $\boldsymbol{\alpha}^{(k)} + d^{(k+1)}/2$ and the algorithm continues till a reduction in the objective function observed. Because, in addition to $\boldsymbol{\alpha}$, $\boldsymbol{\delta}_i$ is also estimated in the PNLS step via the Gauss-Newton algorithm therefore, the same step-halving procedure is also considered for that.

Estimation of $\boldsymbol{\theta}$ and σ^2:
In the LME step, the profiled log likelihood is needed. If $\boldsymbol{\theta}$ is an unconstrained set of parameters such that $\Delta = \Delta(\boldsymbol{\theta})$ then based on (3.17) and given $\hat{\boldsymbol{\alpha}}^{(k+1)}$ and $\hat{\boldsymbol{\delta}}_i^{(k+1)}$ from the PNLS step, the likelihood of pseudo data $\hat{\boldsymbol{Z}}^{(k+1)} := (\hat{Z}_1^{(k+1)}, \cdots, \hat{Z}_I^{(k+1)})$ at the $(k+1)$th iteration is given by

$$L(\boldsymbol{\alpha}, \boldsymbol{\theta}, \sigma^2 | \hat{\boldsymbol{z}}^{(k+1)})$$

$$\approx \prod_{i=1}^{I} \int \frac{\exp[-(||\hat{z}_i^{(k+1)} - \tilde{\boldsymbol{X}}_i^{(k+1)}(\boldsymbol{\alpha} + \boldsymbol{\delta}_i)||^2 + ||\Delta(\boldsymbol{\theta})\boldsymbol{\delta}_i||^2)/2\sigma^2]}{(2\pi\sigma^2)^{N_i/2}(2\pi)^{q/2}\sqrt{|\sigma^2(\Delta(\boldsymbol{\theta})^\intercal\Delta(\boldsymbol{\theta}))^{-1}|}} d\delta_i$$

$$= \prod_{i=1}^{I} \frac{\text{abs}|\Delta(\boldsymbol{\theta})|}{(2\pi\sigma^2)^{N_i/2}} \int \frac{\exp[-(||\hat{\tilde{z}}_i^{(k+1)} - \tilde{\boldsymbol{X}}_i^{(k+1)}(\boldsymbol{\alpha}+\boldsymbol{\delta}_i)||^2 + ||\Delta(\boldsymbol{\theta})\boldsymbol{\delta}_i||^2)/2\sigma^2]}{(2\pi\sigma^2)^{q/2}} d\delta_i$$

$$= \prod_{i=1}^{I} \frac{\text{abs}|\Delta(\boldsymbol{\theta})|}{(2\pi\sigma^2)^{N_i/2}} \int \frac{\exp[-(||\hat{\hat{z}}_i^{(k+1)} - \hat{\boldsymbol{X}}_i^{(k+1)}\boldsymbol{\alpha} - \hat{\boldsymbol{W}}_i^{(k+1)}\boldsymbol{\delta}_i||^2)/2\sigma^2]}{(2\pi\sigma^2)^{q/2}} d\delta_i, \quad (7.28)$$

where $\hat{\boldsymbol{z}}^{(k+1)} := (\hat{z}_1^{(k+1)}, \cdots, \hat{z}_I^{(k+1)})$ is the realization of $\hat{\boldsymbol{Z}}^{(k+1)}$ and $\hat{\hat{z}}_i^{(k+1)}$ is the realization of $\hat{\hat{Z}}_i^{(k+1)}$ which is defined as

$$\hat{\hat{Z}}_i^{(k+1)} = \begin{pmatrix} \hat{Z}_i^{(k+1)} \\ \mathbf{0} \end{pmatrix}$$

and

$$\hat{\boldsymbol{X}}_i^{(k+1)} = \begin{pmatrix} \tilde{\boldsymbol{X}}_i^{(k+1)} \\ \mathbf{0} \end{pmatrix}, \quad \hat{\boldsymbol{W}}_i^{(k+1)} = \begin{pmatrix} \tilde{\boldsymbol{X}}_i^{(k+1)} \\ \Delta(\boldsymbol{\theta}) \end{pmatrix}.$$

Let the QR decomposition of $\hat{\boldsymbol{W}}_i^{(k+1)}$ as

$$\hat{\boldsymbol{W}}_i^{(k+1)} = \tilde{Q}_i \begin{pmatrix} \tilde{R}_{11(i)} \\ \mathbf{0} \end{pmatrix}.$$

Now we apply the same technique in the PNLS step, from (7.20)-(7.25) and replace $\tilde{\hat{Z}}_i^{(k+1)}$, $\hat{\hat{\boldsymbol{W}}}_i^{(k+1)}$, $\hat{\hat{\boldsymbol{X}}}_i^{(k+1)}$, $c_{1(i)}$, $R_{10(i)}$, $R_{11(i)}$, $c_{0(i)}$ and $R_{00(i)}$ by the corresponding values $\hat{\hat{Z}}_i^{(k+1)}$, $\hat{\boldsymbol{W}}_i^{(k+1)}$, $\hat{\boldsymbol{X}}_i^{(k+1)}$, $\tilde{c}_{1(i)}$, $\tilde{R}_{10(i)}$, $\tilde{R}_{11(i)}$, $\tilde{c}_{0(i)}$ and $\tilde{R}_{00(i)}$, where

$$\begin{pmatrix} \tilde{R}_{10(i)} \\ \tilde{R}_{00(i)} \end{pmatrix} = \tilde{Q}_i^{\intercal} \hat{\boldsymbol{X}}_i^{(k+1)}, \quad \begin{pmatrix} \tilde{c}_{1(i)} \\ \tilde{c}_{0(i)} \end{pmatrix} = \tilde{Q}_i^{\intercal} \hat{\hat{Z}}_i^{(k+1)}.$$

The integral term in (7.28) by use of (7.21) (by replacing the mentioned values) can be rewritten as

$$\int \frac{\exp[-(||\hat{\hat{z}}_i^{(k+1)} - \hat{\boldsymbol{X}}_i^{(k+1)}\boldsymbol{\alpha} - \hat{\boldsymbol{W}}_i^{(k+1)}\boldsymbol{\delta}_i||^2)/2\sigma^2]}{(2\pi\sigma^2)^{q/2}} d\delta_i$$

$$= \exp[\frac{||\tilde{c}_{0(i)} - \tilde{R}_{00(i)}\boldsymbol{\alpha}||^2}{-2\sigma^2}] \int \frac{\exp[\frac{||\tilde{c}_{1(i)} - \tilde{R}_{10(i)}\boldsymbol{\alpha} - \tilde{R}_{11(i)}\boldsymbol{\delta}_i||^2}{-2\sigma^2}]}{(2\pi\sigma^2)^{q/2}} d\delta_i. \quad (7.29)$$

Since $\tilde{R}_{11(i)}$ is nonsingular, it is possible to apply a change of variable to $\phi_i = (\tilde{c}_{1(i)} - \tilde{R}_{10(i)}\boldsymbol{\alpha} - \tilde{R}_{11(i)}\boldsymbol{\delta}_i)/\sigma$ with derivative $d\phi_i = \sigma^{-q}\text{abs}|\tilde{R}_{11(i)}|d\boldsymbol{\delta}_i$. Then the integral term in (7.29) can be written as

$$\int \frac{\exp[\frac{||\tilde{c}_{1(i)} - \tilde{R}_{10(i)}\boldsymbol{\alpha} - \tilde{R}_{11(i)}\boldsymbol{\delta}_i||^2}{-2\sigma^2}]}{(2\pi\sigma^2)^{q/2}} d\delta_i = \frac{1}{\text{abs}|\tilde{R}_{11(i)}|} \int \frac{\exp(-||\phi_i||^2/2)}{(2\pi)^{q/2}} d\phi_i$$

$$= \frac{1}{\text{abs}|\tilde{R}_{11(i)}|}. \tag{7.30}$$

Since $\tilde{R}_{11(i)}$ is a triangular matrix, its determinant is readily obtained by product of its diagonal elements. Replacing (7.30) to (7.29) to (7.28) simplify the likelihood expression as

$$L(\boldsymbol{\alpha}, \boldsymbol{\theta}, \sigma^2 | \hat{\boldsymbol{z}}^{(k+1)}) \approx \prod_{i=1}^{I} \frac{\exp[-||\tilde{c}_{0(i)} - \tilde{R}_{00(i)}\boldsymbol{\alpha}||^2/2\sigma^2]}{(2\pi\sigma^2)^{N_i/2}} \text{abs}(\frac{|\Delta(\boldsymbol{\theta})|}{|\tilde{R}_{11(i)}|})$$

$$= \frac{\exp[-\sum_{i=1}^{I} ||\tilde{c}_{0(i)} - \tilde{R}_{00(i)}\boldsymbol{\alpha}||^2/2\sigma^2]}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{I} \text{abs}(\frac{|\Delta(\boldsymbol{\theta})|}{|\tilde{R}_{11(i)}|}).$$

Consider the following QR decomposition

$$\begin{pmatrix} \tilde{R}_{00(1)} & \tilde{c}_{0(1)} \\ \vdots & \vdots \\ \tilde{R}_{00(I)} & \tilde{c}_{0(I)} \end{pmatrix} = \tilde{Q}_0 \begin{pmatrix} \tilde{R}_{00} & \tilde{c}_0 \\ \mathbf{0} & \tilde{c}_{-1} \end{pmatrix}.$$

Now by applying the same technique in (7.25) (by replacing the mentioned values) and use of the last two terms of it, we have

$$L(\boldsymbol{\alpha}, \boldsymbol{\theta}, \sigma^2 | \hat{\boldsymbol{z}}^{(k+1)})$$

$$\approx (2\pi\sigma^2)^{-N/2} \exp(\frac{||\tilde{c}_{-1}||^2 + ||\tilde{c}_0 - \tilde{R}_{00}\boldsymbol{\alpha}||^2}{-2\sigma^2}) \prod_{i=1}^{I} \text{abs}(\frac{|\Delta(\boldsymbol{\theta})|}{|\tilde{R}_{11(i)}|}). \tag{7.31}$$

Now the estimates of $\boldsymbol{\alpha}$ and σ^2 at the $(k+1)$th iteration which maximize the above expression are given by

$$\hat{\tilde{\boldsymbol{\alpha}}}^{(k+1)} = \tilde{R}_{00}^{-1}\tilde{c}_0, \quad \hat{\sigma}^{2(k+1)} = \frac{||\tilde{c}_{-1}||^2}{N}. \tag{7.32}$$

The simplified profile likelihood by given $\hat{\sigma}^{2(k+1)}$ and $\hat{\tilde{\boldsymbol{\alpha}}}^{(k+1)}$ is obtained as follows:

$$L(\boldsymbol{\theta} | \hat{\boldsymbol{z}}^{(k+1)}) = L(\hat{\tilde{\boldsymbol{\alpha}}}^{(k+1)}, \boldsymbol{\theta}, \hat{\sigma}^{2(k+1)} | \hat{\boldsymbol{z}}^{(k+1)})$$

$$\approx (\frac{N}{2\pi||\tilde{c}_{-1}||^2})^{N/2} \exp(-\frac{N}{2}) \prod_{i=1}^{I} \text{abs}(\frac{|\Delta(\boldsymbol{\theta})|}{|\tilde{R}_{11(i)}|}),$$

or the profiled log likelihood as

$$l(\boldsymbol{\theta} | \hat{\boldsymbol{z}}^{(k+1)}) = \log L(\boldsymbol{\theta} | \hat{\boldsymbol{z}}^{(k+1)})$$

$$\approx \frac{N}{2}[\log(N) - \log(2\pi) - 1] - N \log ||\tilde{c}_{-1}|| + \sum_{i=1}^{I} \log \text{abs}(\frac{|\Delta(\boldsymbol{\theta})|}{|\tilde{R}_{11(i)}|}). \tag{7.33}$$

For getting the maximum likelihood estimate $\hat{\boldsymbol{\theta}}^{(k+1)}$, the profiled log likelihood (7.33) is maximized with respect to $\boldsymbol{\theta}$. The algorithm iterates between the PNLS and LME steps till convergence. In this algorithm, we have two estimates for $\boldsymbol{\alpha}$ at each iteration in the PNLS step and LME step (see equations (7.26) and (7.32)). Demidenko (2004) proves that, at convergence, the both procedures i.e., PNLS and LME provide the same estimate of $\boldsymbol{\alpha}$.

7.3 The EM algorithm

The expectation and maximization (EM) algorithm is a widely applicable algorithm in the incomplete data problems which presents an iterative procedure to compute the ML estimations in situations where, algorithms like the Newton-Raphson method may yield to be more difficult. Let $\boldsymbol{Y} := (Y_1^\intercal, \cdots, Y_I^\intercal)^\intercal$ and $\boldsymbol{X} := (X_1^\intercal, \cdots, X_I^\intercal)^\intercal$ be respectively the observed and missing data matrices and $\boldsymbol{y}$ and $\boldsymbol{x}$ be the realizations of $\boldsymbol{Y}$ and $\boldsymbol{X}$. Consider $\boldsymbol{\beta}$ the unknown parameters vector with the parameter space Ω and $\boldsymbol{Z} := (\boldsymbol{Y}^\intercal, \boldsymbol{X}^\intercal)^\intercal$, with the realization $\boldsymbol{z}$, the complete data matrix. The EM algorithm has two steps, the expectation step and the maximization step. At the E step, the conditional expectation of the complete data log likelihood function $\log f(\boldsymbol{z}|\boldsymbol{\beta})$ given the observed data and the current estimate of $\boldsymbol{\beta}$ is calculated. And then at the M step, the maximization of the obtained conditional expectation with respect to $\boldsymbol{\beta}$ over the parameter space Ω is required. At the $(k+1)$th iteration, the EM algorithm is defined as follows:

E-Step. Calculate $Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)})$, where

$$Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}) = \mathbb{E}(\log f(\boldsymbol{Z}|\boldsymbol{\beta})|\boldsymbol{y}, \boldsymbol{\beta}^{(k)}).$$

M-Step. Maximization of $Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)})$ with respect to $\boldsymbol{\beta} \in \Omega$; that is, choose of $\boldsymbol{\beta}^{(k+1)}$ such that

$$\forall \boldsymbol{\beta} \in \Omega, \quad Q(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\beta}^{(k)}) \geq Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}).$$

The E and M step are repeated till $|L(\boldsymbol{\beta}^{(k+1)}) - L(\boldsymbol{\beta}^{(k)})| < \epsilon$, where ϵ is a predetermined small value and $L(\boldsymbol{\beta}^{(k)}) := f(\boldsymbol{y}|\boldsymbol{\beta}^{(k)})$ is the incomplete-data likelihood function.

Lemma 7.1. *Monotonicity of the EM algorithm [Dempster et al., 1977]:*
After an EM iteration, the incomplete-data likelihood function $L(\boldsymbol{\beta})$ is not decreased; that is

$$L(\boldsymbol{\beta}^{(k+1)}) \geq L(\boldsymbol{\beta}^{(k)}). \tag{7.34}$$

Proof. : Let $f(\boldsymbol{z}|\boldsymbol{\beta})$ and $f(\boldsymbol{y}|\boldsymbol{\beta})$ be the probability density functions of $\boldsymbol{Z}$ and $\boldsymbol{Y}$. The conditional density of $\boldsymbol{Z}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is given by

$$f(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\beta}) = f(\boldsymbol{z}|\boldsymbol{\beta})/f(\boldsymbol{y}|\boldsymbol{\beta}), \tag{7.35}$$

Then the log likelihood is calculated as follows

$$\log L(\boldsymbol{\beta}) = \log f(\boldsymbol{y}|\boldsymbol{\beta}) = \log f(\boldsymbol{z}|\boldsymbol{\beta}) - \log f(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\beta}). \tag{7.36}$$

By taking the expectation of both sides of (7.36) with respect to the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{Y} = \boldsymbol{y}$, we have

$$\begin{aligned} \log L(\boldsymbol{\beta}) &= \mathbb{E}(\log f(\boldsymbol{Z}|\boldsymbol{\beta})|\boldsymbol{y}, \boldsymbol{\beta}^{(k)}) - \mathbb{E}(\log f(\boldsymbol{Z}|\boldsymbol{y}, \boldsymbol{\beta})|\boldsymbol{y}, \boldsymbol{\beta}^{(k)}) \\ &= Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}) - H(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}), \end{aligned} \tag{7.37}$$

where

$$H(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}) := \mathbb{E}(\log f(\boldsymbol{Z}|\boldsymbol{y}, \boldsymbol{\beta})|\boldsymbol{y}, \boldsymbol{\beta}^{(k)}). \tag{7.38}$$

From (7.37), we have

$$\begin{aligned} &\log L(\boldsymbol{\beta}^{(k+1)}) - \log L(\boldsymbol{\beta}^{(k)}) \\ &= [Q(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\beta}^{(k)}) - Q(\boldsymbol{\beta}^{(k)}, \boldsymbol{\beta}^{(k)})] - [H(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\beta}^{(k)}) - H(\boldsymbol{\beta}^{(k)}, \boldsymbol{\beta}^{(k)})]. \end{aligned} \tag{7.39}$$

Since $\boldsymbol{\beta}^{(k+1)}$ is selected such that

$$\forall \boldsymbol{\beta} \in \Omega, \quad Q(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\beta}^{(k)}) \geq Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}), \tag{7.40}$$

hence the first difference on the right hand of (7.39) is nonnegative. If the second difference term in (7.39) is nonpositive; that is

$$H(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}) - H(\boldsymbol{\beta}^{(k)}, \boldsymbol{\beta}^{(k)}) \leq 0,$$

then (7.34) holds. For any $\boldsymbol{\beta}$,

$$\begin{aligned} H(\boldsymbol{\beta}, \boldsymbol{\beta}^{(k)}) - H(\boldsymbol{\beta}^{(k)}, \boldsymbol{\beta}^{(k)}) &= \mathbb{E}[\log\{f(\boldsymbol{Z}|\boldsymbol{y}, \boldsymbol{\beta})/f(\boldsymbol{Z}|\boldsymbol{y}, \boldsymbol{\beta}^{(k)})\}|\boldsymbol{y}, \boldsymbol{\beta}^{(k)}] \\ &\leq \log[\mathbb{E}\{f(\boldsymbol{Z}|\boldsymbol{y}, \boldsymbol{\beta})/f(\boldsymbol{Z}|\boldsymbol{y}, \boldsymbol{\beta}^{(k)})\}|\boldsymbol{y}, \boldsymbol{\beta}^{(k)}] \qquad (7.41) \\ &= \log \int f(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\beta}) d\boldsymbol{z} \\ &= 0, \qquad (7.42) \end{aligned}$$

where the inequality in (7.41) holds by use of the Jensen's inequality and the concavity property of the logarithmic function. Finally, based on (7.40) and (7.42) the inequality (7.34), which shows that the incomplete-data likelihood function is not decreased after an EM iteration, holds (McLachlan and Krishnan, 2008). □

7.4 The Newton-Raphson algorithm

Let $\boldsymbol{\beta}$ be the parameters vector in a linear mixed effects model. The Newton-Raphson algorithm uses a first order expansion of the gradiant of the log-likelihood function (score function) around the current estimate $\boldsymbol{\beta}^{(k)}$ to obtain the next estimate $\boldsymbol{\beta}^{(k+1)}$. In each Newton-Raphson iteration, the computation of the score function and its derivative are required. Under general conditions, which usually hold in practice, the Newton-Raphson algorithm is quadratically convergent (Thisted, 1988).

7.5 The calculation of conditional expectations in the Liski and Nummi's estimation method

For the estimation of parameters, the calculation of expectations in (4.15)-(4.17) with respect to $f(y_0^F|y_0^P, \boldsymbol{\beta}^{(k)})$ is needed. For this, we have

$$\mathbb{E}\{\begin{pmatrix} Y_0^P \\ Y_0^F \end{pmatrix} |y_0^P, \boldsymbol{\beta}^{(k)}\} = \begin{pmatrix} y_0^P \\ \hat{y}_0^{F^{(k+1)}} \end{pmatrix},$$

where

$$\begin{aligned}
\hat{y}_0^{F^{(k+1)}} &:= \mathbb{E}(Y_0^F|Y_0^P = y_0^P, \boldsymbol{\beta}^{(k)}) \\
&= \mathbb{E}(Y_0^F|\boldsymbol{\beta}^{(k)}) + \mathbb{C}\text{ov}(Y_0^F, Y_0^P|\boldsymbol{\beta}^{(k)})\mathbb{C}\text{ov}(Y_0^P|\boldsymbol{\beta}^{(k)})^{-1}(y_0^P - \mathbb{E}(Y_0^P|\boldsymbol{\beta}^{(k)})) \\
&= \boldsymbol{x}_0^F\boldsymbol{\alpha}^{(k)} + H_0^{FP}(\boldsymbol{\theta}^{(k)})H_0^{P^{-1}}(\boldsymbol{\theta}^{(k)})(y_0^P - X_0^P\boldsymbol{\alpha}^{(k)}),
\end{aligned}$$

with

$$H_0^P(\boldsymbol{\theta}^{(k)}) := \boldsymbol{X}_0^P D(\boldsymbol{\theta}^{(k)})\boldsymbol{X}_0^{P\intercal} + \boldsymbol{I}_{N_0}, \quad H_0^{FP}(\boldsymbol{\theta}^{(k)}) := \boldsymbol{x}_0^F D(\boldsymbol{\theta}^{(k)})\boldsymbol{X}_0^{P\intercal}.$$

The conditional expectation $\mathbb{E}\{\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}^\intercal H_0^{-1}(\boldsymbol{\theta}) \begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix} |y_0^P, \boldsymbol{\beta}^{(k)}\}$ is calculated as

$$\mathbb{E}\{\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}^\intercal H_0^{-1}(\boldsymbol{\theta}) \begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix} |y_0^P, \boldsymbol{\beta}^{(k)}\}$$

$$= \mathbb{E}\{\text{tr}[\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}^\intercal H_0^{-1}(\boldsymbol{\theta}) \begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}]|y_0^P, \boldsymbol{\beta}^{(k)}\}$$

$$= \mathbb{E}\{\text{tr}[\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix} \begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}^\intercal H_0^{-1}(\boldsymbol{\theta})]|y_0^P, \boldsymbol{\beta}^{(k)}\}$$

$$= \mathrm{tr}[\mathbb{E}\{\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}^\intercal | y_0^P, \boldsymbol{\beta}^{(k)}\} H_0^{-1}(\boldsymbol{\theta}^{(k)})]$$

$$= \mathrm{tr}\{[\begin{pmatrix} y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha}^{(k)} \\ \hat{y}_0^{F(k+1)} - \boldsymbol{x}_0^F\boldsymbol{\alpha}^{(k)} \end{pmatrix}\begin{pmatrix} y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha}^{(k)} \\ \hat{y}_0^{F(k+1)} - \boldsymbol{x}_0^F\boldsymbol{\alpha}^{(k)} \end{pmatrix}^\intercal + \begin{pmatrix} \mathbf{0}_{N_0\times N_0} & \mathbf{0}_{N_0\times 1} \\ \mathbf{0}_{1\times N_0} & V_{FP}(\boldsymbol{\theta}^{(k)}) \end{pmatrix}] H_0^{-1}(\boldsymbol{\theta}^{(k)})\},$$

where y_0^P is the realization of Y_0^P and

$$V_{FP}(\boldsymbol{\theta}^{(k)}) := \mathbb{V}\mathrm{ar}(Y_0^F | y_0^P, \boldsymbol{\beta}^{(k)})$$

$$= \mathbb{V}\mathrm{ar}(Y_0^F|\boldsymbol{\beta}^{(k)}) - \mathbb{C}\mathrm{ov}(Y_0^F, Y_0^P|\boldsymbol{\beta}^{(k)})\mathbb{C}\mathrm{ov}(Y_0^P|\boldsymbol{\beta}^{(k)})^{-1}\mathbb{C}\mathrm{ov}(Y_0^F, Y_0^P|\boldsymbol{\beta}^{(k)})^\intercal$$

$$= \sigma^{2(k)}[H_0^F(\boldsymbol{\theta}^{(k)}) - H_0^{FP}(\boldsymbol{\theta}^{(k)})H_0^{P^{-1}}(\boldsymbol{\theta}^{(k)})H_0^{FP^\intercal}(\boldsymbol{\theta}^{(k)})],$$

with

$$H_0^F(\boldsymbol{\theta}^{(k)}) := \boldsymbol{x}_0^F D(\boldsymbol{\theta}^{(k)})\boldsymbol{x}_0^{F^\intercal} + 1.$$

The conditional expectation in (4.17) can also be calculated by the same way as follows:

$$\mathbb{E}[\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}^\intercal H_0^{-1}(\boldsymbol{\theta})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_w}H_0^{-1}\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix} | y_0^P, \boldsymbol{\beta}^{(k)}]$$

$$= \mathbb{E}\{\mathrm{tr}[\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}^\intercal H_0^{-1}(\boldsymbol{\theta})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_w}H_0^{-1}(\boldsymbol{\theta})\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}]|y_0^P, \boldsymbol{\beta}^{(k)}\}$$

$$= \mathrm{tr}[\mathbb{E}\{\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}\begin{pmatrix} Y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F - \boldsymbol{x}_0^F\boldsymbol{\alpha} \end{pmatrix}^\intercal H_0^{-1}(\boldsymbol{\theta})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_w}H_0^{-1}(\boldsymbol{\theta})|y_0^P, \boldsymbol{\beta}^{(k)}\}]$$

$$= \mathrm{tr}\{[\begin{pmatrix} y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha}^{(k)} \\ \hat{y}_0^{F(k+1)} - \boldsymbol{x}_0^F\boldsymbol{\alpha}^{(k)} \end{pmatrix}\begin{pmatrix} y_0^P - \boldsymbol{X}_0^P\boldsymbol{\alpha}^{(k)} \\ \hat{y}_0^{F(k+1)} - \boldsymbol{x}_0^F\boldsymbol{\alpha}^{(k)} \end{pmatrix}^\intercal$$

$$+ \begin{pmatrix} \mathbf{0}_{N_0\times N_0} & \mathbf{0}_{N_0\times 1} \\ \mathbf{0}_{1\times N_0} & V_{FP}(\boldsymbol{\theta}^{(k)}) \end{pmatrix}] H_0^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_w}H_0^{-1}(\boldsymbol{\theta}^{(k)})\}.$$

It should be noted that for the estimation of $\boldsymbol{\alpha}$ and σ^2 we have the closed form expressions. But for the estimation of $\boldsymbol{\theta}$, there is no closed form expression. Hence, for getting the update estimation of $\boldsymbol{\theta}$ at each EM iteration we apply the following Newton-Raphson method (Gumedze and Dunne, 2011):

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \boldsymbol{\theta}^{(k)} - [\mathbb{E}(\frac{\partial^2 \log(f(\boldsymbol{Z}|\boldsymbol{\beta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal}|\boldsymbol{y}_{obs}, \boldsymbol{\beta}^{(k)})]^{-1}\mathbb{E}(\frac{\partial \log(f(\boldsymbol{Z}|\boldsymbol{\beta}))}{\partial\boldsymbol{\theta}}|\boldsymbol{y}_{obs}, \boldsymbol{\beta}^{(k)}),$$

where

$$\mathbb{E}\{\frac{\partial}{\partial\theta_w}\log f(\boldsymbol{Z}|\boldsymbol{\beta})|\boldsymbol{y}_{obs},\boldsymbol{\beta}^{(k)}\} = -\frac{1}{2}\sum_{i=0}^{I}\mathrm{tr}(H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_w})$$

$$+\frac{1}{2\sigma^{2(k)}}\sum_{i=1}^{I}(y_i-\boldsymbol{X}_i\boldsymbol{\alpha}^{(k)})^{\intercal}H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_w}H_i^{-1}(\boldsymbol{\theta}^{(k)})(y_i-\boldsymbol{X}_i\boldsymbol{\alpha}^{(k)})$$

$$+\frac{1}{2\sigma^{2(k)}}\mathrm{tr}\{[\begin{pmatrix} y_0^P-\boldsymbol{X}_0^P\boldsymbol{\alpha}^{(k)} \\ \hat{y}_0^{F(k+1)}-\boldsymbol{x}_0^F\boldsymbol{\alpha}^{(k)}\end{pmatrix}\begin{pmatrix} y_0^P-\boldsymbol{X}_0^P\boldsymbol{\alpha}^{(k)} \\ \hat{y}_0^{F(k+1)}-\boldsymbol{x}_0^F\boldsymbol{\alpha}^{(k)}\end{pmatrix}^{\intercal}$$

$$+\begin{pmatrix}\mathbf{0}_{N_0\times N_0} & \mathbf{0}_{N_0\times 1} \\ \mathbf{0}_{1\times N_0} & V_{FP}(\boldsymbol{\theta}^{(k)})\end{pmatrix}]H_0^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_w}H_0^{-1}(\boldsymbol{\theta}^{(k)})\},$$

and for $t,w\in\{1,\cdots,s\}$ we have

$$\mathbb{E}(\frac{\partial^2\log(f(\boldsymbol{Z}|\boldsymbol{\beta}))}{\partial\theta_t\partial\theta_w}|y_{obs},\boldsymbol{\beta}^{(k)}) = \frac{1}{2}\sum_{i=0}^{I}\mathrm{tr}(H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_t}H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_w})$$

$$-\frac{1}{2\sigma^{2(k)}}\sum_{i=1}^{I}(y_i-\boldsymbol{X}_i\boldsymbol{\alpha}^{(k)})^{\intercal}[H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_t}H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_w}H_i^{-1}(\boldsymbol{\theta}^{(k)})$$

$$+H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_w}H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_t}H_i^{-1}(\boldsymbol{\theta}^{(k)})](y_i-\boldsymbol{X}_i\boldsymbol{\alpha}^{(k)})$$

$$-\frac{1}{2\sigma^{2(k)}}\mathbb{E}\{\begin{pmatrix}Y_0^P-\boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F-\boldsymbol{x}_0^F\boldsymbol{\alpha}\end{pmatrix}^{\intercal}[H_0^{-1}(\boldsymbol{\theta})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_t}H_0^{-1}(\boldsymbol{\theta})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_w}H_0^{-1}(\boldsymbol{\theta})$$

$$+H_0^{-1}(\boldsymbol{\theta})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_w}H_0^{-1}(\boldsymbol{\theta})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_t}H_0^{-1}(\boldsymbol{\theta})]\begin{pmatrix}Y_0^P-\boldsymbol{X}_0^P\boldsymbol{\alpha} \\ Y_0^F-\boldsymbol{x}_0^F\boldsymbol{\alpha}\end{pmatrix}|y_0^P,\boldsymbol{\beta}^{(k)}\}$$

$$=\frac{1}{2}\sum_{i=0}^{I}\mathrm{tr}(H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_t}H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_w})$$

$$-\frac{1}{2\sigma^{2(k)}}\sum_{i=1}^{I}(y_i-\boldsymbol{X}_i\boldsymbol{\alpha}^{(k)})^{\intercal}[H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_t}H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_w}H_i^{-1}(\boldsymbol{\theta}^{(k)})$$

$$+H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_w}H_i^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_i(\boldsymbol{\theta})}{\partial\theta_t}H_i^{-1}(\boldsymbol{\theta}^{(k)})](y_i-\boldsymbol{X}_i\boldsymbol{\alpha}^{(k)})$$

$$-\frac{1}{2\sigma^{2(k)}}\mathrm{tr}\{[\begin{pmatrix} y_0^P-\boldsymbol{X}_0^P\boldsymbol{\alpha}^{(k)} \\ \hat{y}_0^{F(k+1)}-\boldsymbol{x}_0^F\boldsymbol{\alpha}^{(k)}\end{pmatrix}\begin{pmatrix} y_0^P-\boldsymbol{X}_0^P\boldsymbol{\alpha}^{(k)} \\ \hat{y}_0^{F(k+1)}-\boldsymbol{x}_0^F\boldsymbol{\alpha}^{(k)}\end{pmatrix}^{\intercal}$$

$$+\begin{pmatrix}\mathbf{0}_{N_0\times N_0} & \mathbf{0}_{N_0\times 1}\\ \mathbf{0}_{1\times N_0} & V_{FP}(\boldsymbol{\theta}^{(k)})\end{pmatrix}][H_0^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_t}H_0^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_w}H_0^{-1}(\boldsymbol{\theta}^{(k)})$$

$$+H_0^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_w}H_0^{-1}(\boldsymbol{\theta}^{(k)})\frac{\partial H_0(\boldsymbol{\theta})}{\partial\theta_t}H_0^{-1}(\boldsymbol{\theta}^{(k)})]\}.$$

At the final iteration, $\hat{\Sigma}$ is given by $\hat{\Sigma}=\hat{\sigma}^2\hat{D}$, where $\hat{D}=D(\hat{\boldsymbol{\theta}})$.

7.6 Vec and vech functions and Duplication matrix

The vec function is applied for representation of a matrix M as a vector by stacking its vector columns. Consider M be a $k\times m$ matrix, then $\text{vec}(M)=(M_{11},M_{21},\cdots,M_{k1},\ M_{12},\cdots,M_{km})^\intercal$ is a $km\times 1$ vector. The vech function is used for a $k\times k$ symmetric matrix M by stacking elements of the matrix beginning from the main diagonal, hence the $k(k+1)/2\times 1$ vector vech(M) involves unique elements of matrix M as $\text{vech}(M)=(M_{11},M_{21},\cdots,M_{k1},M_{22},\cdots,M_{kk})^\intercal$. These two functions are relevant via the $k^2\times k(k+1)/2$ duplication matrix $\mathfrak{D}_k$. For each $k\times k$ symmetric matrix M, we have $\text{vec}(M)=\mathfrak{D}_k\text{vech}(M)$ and consequently $\text{vech}(M)=\mathfrak{D}_k^+\text{vec}(M)$, where $\mathfrak{D}_k^+=(\mathfrak{D}_k^\intercal\mathfrak{D}_k)^{-1}\mathfrak{D}_k^\intercal$.

7.7 Block matrix inversion

Assume square matrix $M_{(n+m)\times(n+m)}$ is partitioned as follows:

$$\begin{pmatrix}A_{n\times n} & B_{n\times m}\\ C_{m\times n} & D_{m\times m}\end{pmatrix}.$$

Matrices A and D must be square so that they can be inverted. Moreover, $D-CA^{-1}B$ must be nonsingular. The inverse of M is given by

$$M^{-1}=\begin{pmatrix}A^{-1}+A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1}\\ -(D-CA^{-1}B)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1}\end{pmatrix},$$

or equivalently,

$$M^{-1}=\begin{pmatrix}(A-BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1}\\ -D^{-1}C(A-BD^{-1}C)^{-1} & D^{-1}+D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1}\end{pmatrix}.$$

7.8 Monte Carlo Integration

Suppose that we are interested to compute a complex integral

$$\int h(x)f(x)dx=\mathbb{E}(h(X)),$$

where $f(x)$ is the probability density function of X defined on $\mathbb{R}$ and $h(x)$ is an integrable function of x (the realization of X). If we draw a random sample $x_1, \cdots, x_M$ from $f(x)$ then the Monte Carlo estimation of $\mathbb{E}(h(X))$ is given by

$$\widehat{\mathbb{E}(h(X))} \approx \frac{1}{M} \sum_{m=1}^{M} h(x_m),$$

where M is called the Monte Carlo sample size. This estimate of the integral is a random variable with variance

$$\mathbb{V}\text{ar}(\widehat{\mathbb{E}(h(X))}) = \frac{1}{M} \int (h(x) - \widehat{\mathbb{E}(h(X))})^2 dx = \frac{\sigma^2_{h(x)}}{M},$$

where $\sigma^2_{h(x)}$ can be estimated as

$$\hat{\sigma}^2_{h(x)} = \frac{1}{M-1} \sum_{m=1}^{M} (h(x_m) - \widehat{\mathbb{E}(h(X))})^2.$$

Hence, the standard error of the Monte Carlo estimation is given by

$$\sqrt{\widehat{\mathbb{V}\text{ar}}(\widehat{\mathbb{E}(h(X))})} = \sqrt{\frac{\hat{\sigma}^2_{h(x)}}{M}} = \sqrt{\frac{1}{M(M-1)} \sum_{m=1}^{M} (h(x_m) - \widehat{\mathbb{E}(h(X))})^2}.$$

7.9 Sampling methods

Sometimes we deal with distributions that directly simulating from them is difficult or even impossible. In addition, in some cases, representation of the distribution in an applicable form is not possible. Therefore, we require to apply some sampling methods in which only knowing the functional form of the density f up to a constant is needed. In these methods, an easy to sample density g is proposed which is called the instrumental density. And the density f is called the target density.

7.9.1 Rejection sampling

Let f be the target density, for the rejection sampling method we need to determine the instrumental density g and a constant K such that

$$f(x) \leq Kg(x)$$

on the support of f.

Algorithm 3 Rejection sampling algorithm

1: Generate x from g and u from $\mathbb{U}[0,1]$.
2: Accept $y = x$ if $u \leq f(x)/Kg(x)$.
3: Return to 1. otherwise.

7.9.2 Gibbs sampling

Assume that for $p > 1$, the random variable X can be represented as $X = (X_1, \cdots, X_p)$. In addition, assume that simulation from the univariate conditional densities $f_1, \cdots, f_p$ is possible, where

$$f_i := f(x_i|x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p), \quad i \in \{1, \cdots, p\}.$$

The Gibbs sampling algorithm is started with the initial value $x^{(0)} = (x_1^{(0)}, \cdots, x_p^{(0)})$ and then at the (k+1)th iteration the following algorithm is applied.

The Gibbs sampling algorithm:
1. Draw $x_1^{(k+1)}$ from $f(x_1|x_2^{(k)}, \cdots, x_p^{(k)})$.
2. Draw $x_2^{(k+1)}$ from $f(x_2|x_1^{(k+1)}, x_3^{(k)}, \cdots, x_p^{(k)})$.
$\vdots$
p. Draw $x_p^{(k+1)}$ from $f(x_p|x_1^{(k+1)}, \cdots, x_{p-1}^{(k+1)})$.

The densities $f_1, \cdots, f_p$ are called full conditionals. The advantage of Gibbs sampler is that even in high dimensional problems, all of full conditionals are univariate.

7.9.3 Importance sampling

Geweke (1989) proposes Importance sampling method in which the expectation of the function $\mathbb{E}(h(X))$ is estimated by drawing an independent sample $x_1, \cdots, x_M$ from an importance density $g(x)$ with $g(x) > 0$ and $h(x)f(x) \neq 0$ (whenever $f(x) > 0$). $\mathbb{E}(h(X))$ can be rewritten as

$$\mathbb{E}(h(X)) := \mathbb{E}_{f(x)}(h(X)) = \int h(x)(\frac{f(x)}{g(x)})g(x)dx = \mathbb{E}_{g(x)}[h(X)(\frac{f(X)}{g(X)})]$$

Then the importance sampling estimation of $\mathbb{E}(h(X))$ is given by

$$\widehat{\mathbb{E}(h(X))} = \mathbb{E}_{g(x)}[\widehat{h(X)(\frac{f(X)}{g(X)})}]$$

$$\approx \frac{1}{M}\sum_{m=1}^{M} h(x_m)(\frac{f(x_m)}{g(x_m)})$$
$$= \frac{1}{M}\sum_{m=1}^{M} h(x_m)w_m,$$

where $x_m, \quad m = 1, \cdots, M$ is the random sample from $g(x)$ and $w_m := \frac{f(x_m)}{g(x_m)}$ is the importance weight. The estimated variance of $\widehat{\mathbb{E}(h(X))}$ is given by

$$\widehat{\mathbb{V}\text{ar}}(\widehat{\mathbb{E}(h(X))}) = \frac{1}{M(M-1)}\sum_{m=1}^{M}[h(x_m)w_m - \widehat{\mathbb{E}(h(X))}]^2.$$

Another formulation of importance sampling is

$$\widehat{\widehat{\mathbb{E}(h(X))}} = \frac{\sum_{m=1}^{M} h(x_m)w_m}{\sum_{m=1}^{M} w_m},$$

and the associated variance estimation is

$$\widehat{\mathbb{V}\text{ar}}(\widehat{\widehat{\mathbb{E}(h(X))}}) = \frac{\sum_{m=1}^{M}[h(x_m) - \widehat{\widehat{\mathbb{E}(h(X))}}]^2 w_m^2}{[\sum_{m=1}^{M} w_m]^2}.$$

In use of $\widehat{\widehat{\mathbb{E}(h(X))}}$ we need that the weight function is known up to a constant whereas in use of $\widehat{\mathbb{E}(h(X))}$ the exact form of the weight function is required. If the support of $g(x)$ involves the support of $f(x)$, then it is proved that the importance sampling estimator converges almost surely to $\mathbb{E}(h(X))$ (Geweke, 1989).

7.9.4 Sampling Importance Resampling (SIR)

The Sampling Importance Resampling is a modified form of importance sampling which is related to the rejection sampling technique. In rejection sampling, specification of a suitable value for the constant K for many pairs distributions $f(x)$ and $g(x)$ is impossible. In such cases, the SIR approach is applied. SIR approach has two steps. In the first step, we draw L samples $x_1, \cdots, x_L$ from $g(x)$. Then in the second step, we compute weights $w_1, \cdots, w_L$ as $w_l := \frac{f(x_l)}{g(x_l)}, \quad l = 1, \cdots, L$. Finally, we draw t $(t < L)$ samples from the discrete distribution $x_1, \cdots, x_L$ with probabilites $p_1, \cdots, p_L$, where $p_l := \frac{w_l}{\sum_{l=1}^{L} w_l} \quad l \in \{1, \cdots, L\}$. The SIR algorithm is given by

Algorithm 4 SIR algorithm

1: Simulate L samples from the instrumental density $g(x)$.
2: Calculate importance weights: $w_l := \frac{f(x_l)}{g(x_l)}$ for $l = 1, \cdots, L$.
3: Compute probabilities: $p_l := \frac{w_l}{\sum_{l=1}^{L} w_l}$ for $l \in \{1, \cdots, L\}$.
4: Take t samples from $x_1, \cdots, x_L$ with probabilities $p_1, \cdots, p_L$,.

The obtained t samples are approximately distributed as $f(x)$, however when $L \to \infty$ the probability distribution of t samples tends to the correct distribution (Bishop, 2006).

7.10 Tables

Table 2: Simulation results up to the first 10 observations from the new individual with the near future prediction

Parameters	α_1	α_2	α_3	Σ_{11}	Σ_{22}	Σ_{33}	Σ_{12}	Σ_{13}	Σ_{23}	σ^2
Estimate										
WmSt	36.1	-189	-0.7	16	600	5e-3	18	0.2	2	0.04
mPB	36.2	-189	-0.7	17	702	7e-3	34	0.2	2	0.04
ePB	36.4	-186.5	-0.7	15	700	7e-3	30	0.2	2	0.04
PBLiski	35	-181	-0.7	20	820	0.01	41	0.3	2.5	0.04
PBmSwamy	35	-186	-0.7	17	750.2	8e-3	38	0.2	2	0.04
True Value	36	-190	-0.7	19	846	0.01	48	0.3	2.5	0.04
Bias										
WmSt	0.1	1	0	-3	-246	-4e-3	-30	-0.1	-0.5	0
mPB	0.2	1	0	-2	-144	-3e-3	-14	-0.1	-0.5	0
ePB	0.4	3.5	0	-4	-146	-3e-3	-18	-0.1	-0.5	0
PBLiski	-1	9	0	1	-26	0	-7	0	0	0
PBmSwamy	-1	4	0	-2	-95.8	-2e-3	-10	-0.1	-0.5	0
MSE										
WmSt	0.35	19	2e-4	43	109660	2e-5	1640	3e-2	1.2	4e-5
mPB	0.35	16.2	1e-4	44.2	73052	1e-5	1490	2e-2	1.03	1e-5
ePB	0.6	17	2e-4	46	76012	1e-5	1590	2e-2	1.08	1e-5
PBLiski	2	96	5e-4	13	1450	6e-6	50	8e-4	0.01	5e-7
PBmSwamy	3.2	48.2	7e-4	38	63130	1e-5	1415	2e-2	0.9	5e-7

Table 3: Simulation results up to the first 10 observations from the new individual with the far away future prediction

Parameters	α_1	α_2	α_3	Σ_{11}	Σ_{22}	Σ_{33}	Σ_{12}	Σ_{13}	Σ_{23}	σ^2
Estimate										
WmSt	35.5	-188.5	-0.7	12	530	6e-3	10	0.1	2	0.04
PBLiski	34.4	-182.03	-0.7	18	810	0.01	39	0.3	2.5	0.04
True Value	36	-190	-0.7	19	846	0.01	48	0.3	2.5	0.04
Bias										
WmSt	-0.5	1.5	0	-7	-316	-4e-3	-38	-0.2	-0.5	0
PBLiski	-1.4	7.9	0	2	-36	0	-9	0	0	0
MSE										
WmSt	0.4	18.2	1e-4	58.2	125131	2e-5	2305	6e-2	2	5e-5
PBLiski	2.5	74	2e-4	20	1610	6e-6	84	5e-4	0.01	5e-7

Table 4: Mean of predicted values, biases, MSE's and true values from the simulation study

Method	**WmSt**	**PBSt**	**ePB**	**mPB**
Near future predicted values				
Mean of predictions				
	0.82	0.67	0.57	0.55
Mean of biases				
	0.01	0	0.02	0.1
Mean of MSE's				
	0.06	0.1	8.7	7.6
Mean of true values				
	0.83	0.67	0.6	0.7
Far away future prediction				
Mean of predicted values				
	17.5	17.8	17.8	17.8
Mean of biases				
	0.06	-0.07	-0.4	-0.13
Mean of MSE's				
	1.4	2.5	8.8	9.5
Mean of true values				
	17.6	17.7	17.4	17.7

Table 5: Mean of predicted values, biases, MSE's and true values from the simulation study

Method	**PBLiski**	**PBmSwamy**	**Hall**	**LR**
Near future predicted values				
Mean of predictions				
	0.5	0.71	1.027	0.85
Mean of biases				
	0.002	-0.01	0.002	0.008
Mean of MSE's				
	0.04	0.07	0.05	0.08
Mean of true values				
	0.5	0.7	1.029	0.86
Far away future prediction				
Mean of predicted values				
	17.63	17.85	17.85	17.76
Mean of biases				
	0.2	-0.17	-0.17	0.04
Mean of MSE's				
	0.85	8.4	0.9	2.1
Mean of true values				
	17.44	17.68	17.68	17.72

Table 6: Prediction interval results from the simulation study

Method	**WmSt**	**mStConf**	**PBSt**	**StConf**	**LR**
Near future results					
Average width					
	10.3	19.2	0.4	0.7	1.1
Mean of emperical coverage rates					
	90%	93%	40%	73%	94%
Average interval score					
	15.3	22.9	4.9	2.4	1.45
Far away future results					
Average width					
	10.13	16.9	2.3	6.1	5.6
Mean of emperical coverage rates					
	93%	96%	46%	87%	94%
Average interval score					
	13.18	18.21	20.9	9.4	6.7

Table 7: Prediction interval results from the simulation study

Method	**PBLiski**	**PBmSwamy**	**Hall**	**ePB**	**mPB**
Near future results					
Average width					
	0.92	1.1	0.9	11.3	11.1
Mean of emperical coverage rates					
	96%	96%	95%	94%	93%
Average interval score					
	0.99	1.3	1.05	14.3	14.3
Far away future results					
Average width					
	3.4	10.6	3.4	10.6	10.4
Mean of emperical coverage rates					
	93%	92 %	93%	91%	95%
Average interval score					
	4.36	13.5	4.4	15.04	12.7

Table 8: Parameter estimates with their standard errors in parentheses using the first 131 observations of the new series

Parameters	**WmSt**[1]	**mPB**[2]	**PBLiski**[1]	**PBmSwamy**
α_1	35.02	35.13	35.08	35.06
	(0.43)	(0.48)	(0.1)	(0.48)
α_2	-188.58	-189.02	-189.18	-189.33
	(2.3)	(5.22)	(1.04)	(5.5)
α_3	-0.76	-0.76	-0.76	-0.76
	(0.08)	(0.015)	(0.003)	(0.016)
Σ_{11}	7.07	8.4	7.7	8.02
	(1.8)	(2.6)	(0.14)	(1.9)
Σ_{22}	745.2	944.2	936.4	1047.5
	(123.5)	(419.2)	(17.5)	(257.8)
Σ_{33}	0.007	0.008	0.008	0.009
	(3e-3)	(3e-3)	(1e-4)	(2e-3)
Σ_{12}	25.7	33.3	32.6	34.7
	(34.01)	(23.1)	(0.6)	(17.07)
Σ_{13}	0.2	0.2	0.2	0.2
	(0.07)	(0.07)	(0.003)	(0.06)
Σ_{23}	2.1	2.6	2.6	2.9
	(1.6)	(1.2)	(0.05)	(0.7)
σ^2	0.04	0.04	0.04	0.04
	(6e-4)	(1e-3)	(6e-4)	(6e-4)
logl	1091.7	1184.5	1134.7	1162.8

[1]Estimations based on the near future prediction at $x_0^F = 35.2$

[2]Estimation method of Pinheiro and Bates using all observations

8 Bibliography

[1] Bates, D. M., Watts, D. G., 1988. *Nonlinear Regression Analysis and its Application*. John Wiley & Sons.

[2] Bishop, C. M., 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin.

[3] Bondeson, J., 1990. *Prediction in random coefficient regression models*. Biometrical Journal, 32(4), 387–405.

[4] Booth, G. J., Hobert, P. J., 1999. *Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm*. Journal of the Royal Statistical Society, B 61, 265–285.

[5] Bruckner-Foit, A., Meyer, S., Moslang, A., 2003. *A stochastic simulation model for micro-cracks in martensitic steel*, Computational Materials Science, 26, 102–110.

[6] Carter, R. L., Yang. M. C. K., 1986. *Large sample inference in random coefficient regression models*. Communications in Statistics–Theory and Methods, 15(8), 2507–2525.

[7] Chivers C (2012) MHadaptive: General Markov Chain Monte Carlo for Bayesian Inference using adaptive Metropolis-Hastings sampling. R package version 1.1–8, url: https://CRAN.R-project.org/package=MHadaptive

[8] Davidian, M., Giltinan, D. M., 2003. *Nonlinear models for repeated measurements: An overview and update*. Journal of Agricultural, Biological and Environmental Statistics, 8, 387–419.

[9] Demidenko, E., 2004. *Mixed Models–Theory and Applications*. John Wiley & Sons, New York.

[10] Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Ser. B, 39, 1–22.

[11] Galecki, A. T., Burzykowski, T., 2013. *Linear Mixed Effects Models using R*. Springer, USA.

[12] Geneiting, T., Raftery, A. E., 2007. *Strictly proper scoring rules, prediction, and estimation*. Journal of the American Statistical Association, 102, 359–378.

[13] Geweke, J., 1989. *Bayesian inference in econometric models using Monte Carlo integration*. Econometrics, 57, 1317–1340.

[14] Gumedze, F. N., Dunne, T. T., 2011. *Parameter estimation and inference in the linear mixed model*. Linear Algebra and its Application, 435, 1920–1944.

[15] Hall, D. B., Clutter, M., 2004. *Multivariate multilevel nonlinear mixed effects models for timber yield predictions*. Biometrics, 60, 16–24.

[16] Hermann, S., Ickstadt, K., Müller, Ch. H., 2016. *Bayesian prediction of crack growth based on a hierarchical diffusion model*. Applied Stochastic Models in Business and Industry, 32(4), 494-510.

[17] Heron, E. A., Walsh, C. D., 2008. *A continuous latent spatial model for crack initiation in bone cement*. Journal of Applied Statistics, 57, 25–42.

[18] Ihara, C., Tanaka, T., 2000. *A stochastic damage accumulation model for crack initiation in high-cycle fatigue*. Fatigue & Fracture of Engineering Materials & Structures, 23, 280–375.

[19] Jamshidian, M., Jennrich, R. I., 1993. *Conjugate gradient acceleration of the EM algorithm*. Journal of the American Statistical Association, 88, 221–228.

[20] Jank, W., 2006. *Implementing and diagnosing the stochastic approximation EM algorithm*. Journal of Computational and Graphical Statistics, 15, 803–829.

[21] Jennrich, R. I., 1969. *Asymptotic properties of non-linear least squares estimators*. The Annals of Mathematical Statistics, 40, 633-643.

[22] Johansen, S., 1982. *Asymptotic inference in random coefficient regression models*. Scandinavian Journal of Statistics, 9, 201–207.

[23] Kadiyala, K. R., Oberhelman, D., 1982. *Response predictions in regressions on panel data*. Communications in Statistics–Theory and Methods, 11(23), 2699–2714.

[24] Koch, K. R., 2007. *Gibbs sampler by sampling importance resampling*. Journal of Geodesy, 81, 581–591.

[25] Kuhn, E., Lavielle, M., 2005. *Maximum likelihood estimation in nonlinear mixed effects models*. Computational Statistics & Data Analysis, 49, 1020–1038.

[26] Laird, N. M., Ware, J. H., 1982. *Random effects models for longitudinal data*. Biometrics, 38(4), 963–974.

[27] Lin, T. I., Lee, J. C., 2006. *A robust approach to t linear mixed models applied to multiple sclerosis data*. Statistics in Medicine, 25, 1397–1412.

[28] Lindstrom, M., Bates, D. 1990. *Nonlinear mixed effects models for repeated measures data*. Biometrics, 44, 673–687.

[29] Liski, E. P., Nummi, T., 1996. *Prediction in repeated-measures models with engineering applications*. Technometrics, 38 (1), 25–36.

[30] Louis, T. A., 1982. *Finding the observed information matrix when using the EM algorithm*. Journal of the Royal Statistical Society, B 44, 226–233.

[31] Mathew, T., Menon, S., Perevozskaya, I., Weerahandi, S., 2016. *Improved prediction intervals in heteroscedastic mixed-effects models*. Statistics & Probability Letters, 114, 48–53.

[32] McLachlan, G. J., Krishnan, T., 2008. *The EM algorithm and Extensions*. Willey series in Probability and Statistics, New York.

[33] Meza, C., Osorio, F., Cruz, R. D., 2012. *Estimation in nonlinear mixed-effects models using heavy-tailed distributions*. Statistics and Computing, 22, 121–139.

[34] Müller, Ch. H., Szugat, S., Maurer, R., 2016. *Simulation free prediction intervals for a state dependent failure process using accelerated lifetime experiments*. SFB Discussion Paper 45/2016, Dortmund.

[35] Nie, L., 2007. *Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: Theory and applications*. Journal of Statistical Planning and Inference, 137, 1787–1804.

[36] Pinheiro, J. C., 1994. *Topics in Mixed Effects Models*. Ph.D. thesis, University of Wisconsin, Madison, WI.

[37] Pinheiro, J. C., Bates, D. M., 2000. *Mixed Effects Models in S and S-plus*. Springer-Verlag, New York.

[38] Pinheiro, J. C., Liu, C., Wu, Y., 2001. *Efficient algorithms for robust estimation in linear mixed effects models using the multivariate t distribution*. Journal of Computational and Graphical Statistics, 10, 249–276.

[39] Pinheiro J C, Bates D, DebRoy S, Sarkar D, R Core Team (2016) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1–128, url: https://CRAN.R-project.org/package=nlme

[40] R Development Core Team, 2016. *R: A language and environment for statistical computing. R foundation for statistical computing*, Vienna, Austria, http://cran.r-project.org/.

[41] Rao, C. R., 1965. *The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves.* Biometrika, 52(3), 447–458.

[42] Rao, C. R., 1975. *Simultaneous estimation of parameters in different linear models and applications to biometric problems.* Biometrics, 31(2), 545–554.

[43] Robert, C. P., Casella, G., 1999. *Monte Carlo Statistical Methods.* Springer-Verlag, New York.

[44] Seber, G., Wild, C., 1989. *Nonlinear Regression.* Wiley, New York.

[45] Sheiner, L. B., Beal, S. L., 1980. *Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model:routine clinical pharmacokinetic data.* Journal of Pharmacokinetics and Pharmacodynamics, 8, 553–571.

[46] Sobczyk, K., Spencer, B. F., 1992. *Random Fatigue. From Data to Theory.* Academic Press, London.

[47] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Linde, A. V. D., 2002. *Bayesian measures of model complexity and fit.* Journal of the Royal Statistical Society, Ser. B, 64(4), 583–639.

[48] Staudenmayer, J., Lake, E. E., Wand, M. P., 2009. *Robustness for general design mixed models using the t-distribution.* Statistical Modelling, 9, 235–255.

[49] Stirnemann, J. J., Samson, A., Thalabard, J. C., 2011. *Individual predictions based on nonlinear mixed modeling: application to prenatal twin growth.* Statistics in Medicine, 31, 1986–1999.

[50] Swamy, P. A. V. B., 1971. *Statistical Inference in Random Coefficient Regression Models.* Springer-Verlag, New York.

[51] Thisted, R. A. E., 1988. *Elements of Statistical Computing.* Chapman and Hall, London.

[52] Venables, W. N., Ripley, B. D., 2002. *Modern Applied Statistics with S.* Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

[53] Virkler, D., Hillberry, B., Geol, P., 1979. *The statistical nature of fatigue crack propagation.* Journal of Engineering Materials and Technology, 101, 148–153.

[54] Vonesh, E. F., Carter, R. L., 1992. *Mixed effects nonlinear regression for unbalanced repeated measures.* Biometrics, 48, 1–18.

[55] Vonesh, E. F., Carter, R. L., 1996. *A note on the use of Laplace's approximation for nonlinear mixed effects models.* Biometrika, 83, 447–452.

[56] Walker, S., 1996. *An EM algorithm for nonlinear random effects models.* Biometrics, 52, 934–944.

[57] Wang, J., 2007. *EM algorithms for nonlinear mixed effects models.* Computational Statistics and Data Analysis, 51, 3244–3256.

[58] Wei, G., Tanner, M., 1990. *A Monte Carlo implemention of the EM algorithm and the poor man's data augmentation algorithms.* Journal of the American Statistical Association, 85, 699–704.

[59] Welsh, A. H., Richardson, A. M., 1997. *Approaches to the robust estimation of mixed models.* In: Maddala, G. S., Rao, C. R. (eds) Handbook of Statistics 15, 343–384.

[60] Wolfinger, R. D., 1993. *Laplacian's approximation for nonlinear mixed models.* Biometrika, 80, 791–795.

[61] Wolfinger, R. D., Lin, X., 1997. *Two Taylor-series approximation methods for nonlinear mixed models.* Computational Statistics & Data Analysis, 25, 465–490.

[62] Wu, L., 2004. *Exact and approximate inferences for nonlinear mixed-effects models with missing covariates.* Journal of the American Statistical Association, 99, 700–709.

[63] Wu, L., 2010. *Mixed Effects Models for Complex Data.* Chapman and Hall/CRC, New York.

[64] Yang, Y., Simpson, D., 2010. *Unified computational methods for regression analysis of zero inflated and bound inflated data.* Computational Statistics & Data Analysis, 54(6), 1525–1534.

[65] Yeap, B. Y., Davidian, M., 2001. *Robust two-stage estimation in hierarchical nonlinear models.* Biometrics, 57, 266–272.