




Selection of terms in random coefficient regression models

Francisco M. M. Rocha & Julio M. Singer


To cite this article: Francisco M. M. Rocha & Julio M. Singer (2017): Selection of terms in random coefficient regression models, Journal of Applied Statistics, DOI: [10.1080/02664763.2016.1273884](https://doi.org/10.1080/02664763.2016.1273884)

To link to this article: <http://dx.doi.org/10.1080/02664763.2016.1273884>

 View supplementary material [↗](#)

 Published online: 02 Jan 2017.

 Submit your article to this journal [↗](#)

 View related articles [↗](#)

 View Crossmark data [↗](#)

Selection of terms in random coefficient regression models

Francisco M. M. Rocha^a and Julio M. Singer^b

^aEscola Paulista de Política, Economia e Negócios, Universidade Federal de São Paulo, São Paulo, Brazil;

^bDepartamento de Estatística, Universidade de São Paulo, São Paulo, Brazil

ABSTRACT

The selection of suitable terms in random coefficient regression models is a challenging problem to practitioners. Although many techniques, ranging from those with a theoretical flavour to those with an exploratory spirit, have been proposed for such purposes, no particular one may be considered as a paradigm. In fact, many authors advocate that they should be used in a complementary way. We consider exploratory methods based on fitting standard regression models to the individual response profiles or to the rows of the sample within-units covariance matrix (for balanced data) that may serve as additional tools in the process of selecting an appropriate model. We evaluate the performance of the proposal via a simulation study and consider applications to two examples in the field of Biostatistics.

ARTICLE HISTORY

Received 23 July 2015

Accepted 26 November 2016

KEYWORDS

Covariance matrix; longitudinal data; mixed models; parallel plots; random coefficient models; repeated measures

1. Introduction

Linear mixed models have gained popularity because of their flexibility to represent repeated measures data. In particular, they have been extensively used in longitudinal studies in the form of random coefficient regression models, where the response for different sample units is modelled by specific polynomials.

Since both the estimation of the parameters associated to the fixed coefficients and the prediction of the random effects rely on an appropriate specification of the within-unit covariance structure induced by the latter, their choice is intimately related to the efficiency of estimators and predictors. In this context, a crucial portion of the analysis effort is related to the choice of the fixed and random coefficients to be included in the model.

Many alternatives have been proposed to solve this challenging problem. Pu and Niu [12] suggest the use of the extended generalized information criterion and Orelie and Edwards [10] consider R^2 statistics to select fixed coefficients. Although both criteria may be used quite efficiently for such purposes, the performance of the procedure proposed by Pu and Niu [12] is relatively poor with respect to the selection of random coefficients, as observed by these authors. Other simple analytical tools have been considered in the literature for the selection of both fixed and random terms. Fearn [4], for example, highlights the importance of examining the nature of the individual profiles for the selection of the random coefficients. Simultaneous plots of such profiles, known as *profile plots* or *parallel*

CONTACT Francisco M. M. Rocha ✉ fmmrocha@unifesp.br

 Supplemental data for this article can be accessed at doi:10.1080/02664763.2016.1273884

plots, introduced by Rao and Rao [13] for describing longitudinal data along with a *loess* smoothed average profile, constitute a widely used graphical technique for an initial selection of the terms in a linear mixed model as well as for checking some model assumptions (see, e.g. [20]).

Tests of hypotheses regarding the significance of the variance components have been suggested by Stram and Lee [17]. This, however, is associated with some technical problems related to the fact that the null hypothesis places the parameter on the border of the parametric space, as indicated in [5] among others. Grady and Helms [6] suggest a plot of within-units covariances as a function of the lag between observations to identify possible auto-correlation structures. Other graphical diagnostic tools include residual as well as global or local influence analyses considered in [16].

None of the available alternatives should be used as the only procedure to select the fixed and random coefficients in linear mixed models. In fact they should be taken as complementary and the decision should be based on simultaneous analyses. Our objective is to propose two additional procedures for such purposes.

As in [11,14], we consider the analysis of the individual profiles through simple within-unit regression models as a tool for better understanding the information contained in the profile plots. Rutter and Elashoff [14] propose an evaluation of the within-unit individual regressions via R^2 and via an examination of the profile plots to identify, for example, an heteroskedastic behaviour (fanning), but they do not propose specific tools to select random coefficients. We also build upon the ideas of Suyama [18], who suggested that the rows of the sample covariance matrix behave as the individual profiles. Specifically, we show how to select fixed and random coefficients through an analysis of the estimated individual regression parameters. The data may be collected at irregularly spaced time points, that is, data for different units may be collected at different time points, but the number of repeated measurements for each unit has to be greater than the number of individual parameters. Also, the functional form of the relation between response and explanatory variable should be the same for every unit. When the data are collected at the same time points (not necessarily equally spaced) for all units, we may use the rows of the within-units sample covariance matrix as an additional tool to identify random coefficients to be included in the model.

The random coefficient regression model is described in Section 2. In Section 3 we show how standard regression methods may be employed for the selection of fixed and random coefficients to be included in homoskedastic conditional independence regression models. First we consider the analysis of the individual response profiles and then we show how the analysis of the rows of the sample covariance matrix may be employed for selecting the random coefficients in the case of balanced data. In Section 4 we illustrate the methods by means of simulated examples and in Section 5, we apply the results to examples in the field of Biostatistics. We conclude with a brief discussion in Section 6.

2. The model

The well known linear mixed model introduced by Laird and Ware [8] for the analysis of longitudinal data is

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad (1)$$

where \mathbf{y}_i is a $m_i \times 1$ vector of observations on the i th unit, $i = 1, \dots, n$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ is a $p \times 1$ vector of unknown population parameters (*fixed coefficients*), \mathbf{X}_i is a $m_i \times p$ known specification matrix (or regression design matrix) corresponding to the fixed coefficients, $\mathbf{b}_i = (b_{i1}, \dots, b_{iq})^\top$ is a $q \times 1$ vector of unknown random effects (*random coefficients*), \mathbf{Z}_i is a $m_i \times q$ known specification matrix (or matrix of within-unit regressors) corresponding to the random coefficients and \mathbf{e}_i is an $m_i \times 1$ vector of random error terms. Usual assumptions include independence between the \mathbf{b}_i and the \mathbf{e}_i , $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$, where $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta})$ is a $q \times q$ positive-definite covariance matrix, and $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i)$, where $\mathbf{R}_i = \mathbf{R}_i(\boldsymbol{\theta})$ is a $m_i \times m_i$ positive-definite covariance matrix and $\boldsymbol{\theta}$ is an $r \times 1$ vector of covariance parameters, functionally independent of $\boldsymbol{\beta}$. When the columns of \mathbf{X}_i and \mathbf{Z}_i correspond to values of powers of the explanatory variable, the model is known as random coefficient regression model.

Frequently, one sets $\mathbf{R}_i = \sigma^2 \mathbf{I}_{m_i}$, with \mathbf{I}_r denoting an r -dimensional identity matrix; the resulting model is known as the *homoskedastic conditional independence model* because given the random coefficients \mathbf{b}_i , the corresponding *conditional model* implies independence of the observations on the i th unit. The marginal distribution of the vector of observations from the i th unit has mean vector $\mathbf{X}_i \boldsymbol{\beta}$ and within-unit covariance matrix

$$\mathbb{V}(\mathbf{y}_i) = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i^\top + \sigma^2 \mathbf{I}_{m_i}. \tag{2}$$

The first component of (2) corresponds to the dispersion of the observations of individual response profiles around the mean response profile and the second component corresponds to the conditional within-unit dispersion of the response. The simplest case, where a single random intercept is included, induces a uniform within-unit covariance matrix (i.e. with equal variances and equal covariances). The most complex case, where the maximum number of possible random terms are included in the model, induces an unstructured (i.e. with possibly different variances and covariances) covariance matrix. If on the one hand, the simplicity of the uniform structure may not be adequate to model real data, on the other, the excessive number of parameters associated to the unstructured covariance matrix may jeopardize the efficiency of estimators, specially for small or moderate sample sizes. Models with a small number (e.g. 2, 3 or 4) of random coefficients are thus interesting alternatives to the two extreme cases mentioned above; they relax the strong restrictions on the covariance structure imposed by the single random intercept model without the over-parametrization required by the more general case.

These models has been widely used for analysing longitudinal data not only because of their simplicity, but also because they can accommodate cases where observations are incomplete or collected irregularly along time. For details, the reader is referred to Verbeke and Molenberghs [19], Diggle *et al.* [3] or Demidenko [2], among others. Random coefficient regression models are also attractive because of their straightforward interpretation as pointed by Fearn [4], for example.

3. Selection of fixed and random coefficients

Assuming homoskedastic conditional independence, we observe that (1) may be re-expressed as

$$\mathbf{y}_i = \mathbf{X}_i^* \boldsymbol{\beta}_i^* + \mathbf{e}_i, \tag{3}$$

where \mathbf{X}_i^* is a matrix with p^* columns obtained from the elements of \mathbf{X}_i and \mathbf{Z}_i ; the columns of \mathbf{X}_i^* are those common to \mathbf{X}_i and \mathbf{Z}_i plus those that are unique either to \mathbf{X}_i or to \mathbf{Z}_i . In many practical problems, the columns of \mathbf{Z}_i correspond to a subset of the columns of \mathbf{X}_i . The elements of $\boldsymbol{\beta}_i^*$ are given by $\beta_k + b_{ik}$ if column k is common to \mathbf{X}_i and \mathbf{Z}_i , by β_k if column k is unique to \mathbf{X}_i or by b_{ik} if column k is unique to \mathbf{Z}_i . We can therefore write $\boldsymbol{\beta}_i^* = \boldsymbol{\beta}^* + \mathbf{b}_i^*$, where null elements may be added to the original $\boldsymbol{\beta}$ and \mathbf{b}_i vectors, so that they have the same dimension. For example, letting

$$\mathbf{X}_i = \begin{pmatrix} 1 & x_{i1} \\ 1 & x_{i2} \\ 1 & x_{i3} \\ 1 & x_{i4} \end{pmatrix}, \quad \mathbf{Z}_i = \begin{pmatrix} 1 & x_{i1}^2 \\ 1 & x_{i2}^2 \\ 1 & x_{i3}^2 \\ 1 & x_{i4}^2 \end{pmatrix}, \quad \boldsymbol{\beta} = (\beta_0, \beta_1)^\top, \quad \mathbf{b} = (b_{i0}, b_{i3})^\top,$$

then

$$\boldsymbol{\beta}_i^* = \begin{pmatrix} \beta_0 + b_{i0} \\ \beta_1 \\ b_{i3} \end{pmatrix} \quad \text{and} \quad \mathbf{X}_i^* = \begin{pmatrix} 1 & x_{i1} & x_{i1}^2 \\ 1 & x_{i2} & x_{i2}^2 \\ 1 & x_{i3} & x_{i3}^2 \\ 1 & x_{i4} & x_{i4}^2 \end{pmatrix}, \quad \boldsymbol{\beta}^* = (\beta_0, \beta_1, 0)^\top,$$

$$\mathbf{b}_i^* = (b_{i0}, 0, b_{i3})^\top.$$

Regarding (1) as a two-stage model, it follows that $\mathbf{y}_i | \mathbf{b}_i \sim \mathcal{N}(\mathbf{X}_i^* \boldsymbol{\beta}_i^*; \sigma^2 \mathbf{I}_{m_i})$, so in the first stage we may consider a set of standard regression models for which the estimated parameters, $\hat{\boldsymbol{\beta}}_i^* = (\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1} \mathbf{X}_i^{*\top} \mathbf{y}_i$ are normally distributed with mean $\boldsymbol{\beta}_i^*$ and covariance matrix $\sigma^2 (\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1}$. In the second stage, we assume that $\boldsymbol{\beta}_i^* \sim \mathcal{N}(\boldsymbol{\beta}^*; \mathbf{G}^*)$, where \mathbf{G}^* consists of \mathbf{G} augmented with null rows and/or columns corresponding to null elements in the random vectors \mathbf{b}_i^* so that the marginal distribution of $\hat{\boldsymbol{\beta}}_i^*$ is $\mathcal{N}(\boldsymbol{\beta}^*; \sigma^2 (\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1} + \mathbf{G}^*)$.

3.1. Selection of fixed coefficients based on the individual response profiles

An unbiased estimator of $\boldsymbol{\beta}^*$ obtained from the individual parameter estimates $\hat{\boldsymbol{\beta}}_i^*$ is the sample average

$$\bar{\boldsymbol{\beta}}^* = (\bar{\beta}_1^*, \dots, \bar{\beta}_{p^*}^*)^\top = n^{-1} \sum_{i=1}^n (\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1} \mathbf{X}_i^{*\top} \mathbf{y}_i \quad (4)$$

for which the (unconditional) variance is

$$\mathbb{V}(\bar{\boldsymbol{\beta}}^*) = n^{-2} \sum_{i=1}^n [\sigma^2 (\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1} + \mathbf{G}^*]. \quad (5)$$

To test whether the k th element of $\boldsymbol{\beta}^*$ is zero, we propose the statistic

$$t = \frac{\bar{\beta}_k^*}{n^{-1} \sqrt{\hat{\sigma}^2 \text{diag}_k [(\sum_{i=1}^n \mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1}]}} \quad (6)$$

where $\text{diag}_k(\mathbf{A})$ denotes the k th element of the main diagonal of a square matrix \mathbf{A} and

$$\hat{\sigma}^2 = \sum_{i=1}^n \frac{m_i - p^*}{\nu} \hat{\sigma}_i^2 \tag{7}$$

with

$$\hat{\sigma}_i^2 = \frac{1}{m_i - p^*} \mathbf{y}_i^\top [\mathbf{I}_{m_i} - \mathbf{X}_i^* (\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1} \mathbf{X}_i^{*\top}] \mathbf{y}_i \quad \text{and} \quad \nu = \sum_{i=1}^n m_i - np^*.$$

We refer Equation (6) to the t distribution with ν degrees of freedom, observing that its denominator is expected to be smaller than the square root of the corresponding element in the estimator of (5). This provides a conservative test, in the sense that it will identify more candidates for fixed coefficients than the final model would possibly have, a feature that is recommendable at the model selection stage. By using only ordinary regression results, the analysis maintains the simplicity convenient for exploratory analysis.

3.2. Selection of random coefficients based on the individual response profiles

The variance of $\hat{\beta}_{ik}^*$, $i = 1, \dots, n$ is expected to be equal to the k th diagonal term of $\sigma^2 (\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1}$ when the variance of the corresponding random coefficient, b_{ik} , is null. Otherwise, we might expect a larger variability of the $\hat{\beta}_{ik}^*$ around its mean. Now, under model (3), the k th element of $\hat{\beta}_i^*$, namely, $\hat{\beta}_{ik}^*$, follows a $\mathcal{N}(\beta_{ik}^*; v_{ik}\sigma^2)$ distribution where $v_{ik} = \text{diag}_k\{(\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1}\}$. Therefore $\hat{\beta}_{ik}^*/\sqrt{v_{ik}} \sim \mathcal{N}(\beta_{ik}^*/\sqrt{v_{ik}}; \sigma^2)$. Letting $\hat{w}_{ik} = \hat{\beta}_{ik}^*/\sqrt{v_{ik}}$ and $\bar{w}_k = n^{-1} \sum_{i=1}^n \hat{w}_{ik}$, it follows that

$$t(\hat{w}_k) = \sqrt{n/(n-1)} (\hat{w}_{ik} - \bar{w}_k) / \hat{\sigma} \sim t_\nu. \tag{8}$$

Thus, for each k we expect around $\alpha\%$ of the values of $t(\hat{w}_k)$ outside the corresponding global significance level $\alpha^* = \alpha/(np^*)$ Bonferroni-corrected confidence interval, namely $[t_\nu(\alpha^*/2), t_\nu(1 - \alpha^*/2)]$ where $t_\nu(\delta)$ denotes the $100\delta\%$ percentile of the t distribution with ν degrees of freedom. A larger percentage of points outside that interval suggests that b_{ik} may be a random coefficient.

Although the method we propose should be viewed with an exploratory spirit, a guideline to include a coefficient as a fixed or a random term in the model is outlined in Table 1. The choice of the significance level α is arbitrary, but we recommend setting $\alpha = 1\%$ when the sample size is small, especially for the selection of the random coefficients, given that profile plots are excellent tools for the specification of the fixed coefficients.

Table 1. Guidelines for inclusion of fixed and random coefficients in the linear mixed model.

Significance of $\hat{\beta}_k^*$ as obtained via (6)	Proportion of $\hat{\beta}_{ik}^*$ outside $[t_\nu(\alpha^*/2), t_\nu(1 - \alpha^*/2)]$	
	$\leq \alpha^*$	$> \alpha^*$
YES	Include only β_k as a fixed coefficient	Include β_k as a fixed coefficient and b_{ik} as a random coefficient
NO	Include neither β_k nor b_{ik} in the model	Include only b_{ik} as a random coefficient

3.3. Selection of random coefficients based on the columns of the sample covariance matrix

Under the homoskedastic conditional independence random coefficients model, the covariance structure of the observed values is governed by the random coefficients, as can be deduced from the corresponding marginal covariance matrix (2). The measurement error variance (σ^2) is only related to the marginal variances. Letting $\mathbf{z}_{i_s}^\top$ denote the s th row of \mathbf{Z}_i , it follows that the random coefficients contribution to the s th column of the marginal covariance matrix is given by $\mathbf{Z}_i \mathbf{G} \mathbf{z}_{i_s}$ which has the same form as the individual profiles $\mathbf{Z}_i \mathbf{b}_i$. When the data for all subjects are collected at the same time points, that is, $m_i = m$, we have $\mathbf{Z}_i = \mathbf{Z}$ so that $\mathbf{V} = \mathbf{Z} \mathbf{G} \mathbf{Z}^\top$. For well specified models, \mathbf{V} may be estimated by $\mathbf{S} - \hat{\sigma}^2 \mathbf{I}_m$ where

$$\mathbf{S} = (n - 1)^{-1} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top.$$

We propose to fit polynomial models (of the same degree) to each of the m rows of the sample covariance matrix \mathbf{S} and to analyse the results along the same lines considered for the response profiles with the objective of providing an additional tool for the selection of random coefficients.

4. Simulation studies

To evaluate the performance of the proposed procedures under three different specifications of the underlying distributions and error term covariance structure, we conducted two simulation studies.

In the first study, the data were generated from a linear mixed model with fixed and random intercepts and slopes, that is, from model (1) with

$$\mathbf{X}_i = \mathbf{Z}_i = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}^\top, \quad \boldsymbol{\beta} = (40, 5)^\top, \quad \text{and}$$

$$\mathbf{G} = \begin{bmatrix} \sigma_a^2 & \rho_{ab} \sigma_a \sigma_b \\ \rho_{ab} \sigma_a \sigma_b & \sigma_b^2 \end{bmatrix}, \quad i = 1, \dots, n,$$

under the following distributions

- (A) $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$ and $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_5)$;
 (B) $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$ and $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{R})$, with

$$\mathbf{R} = \frac{1}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}, \quad 0 < \rho < 1;$$

- (C) $\mathbf{b}_i \sim t_4(\mathbf{0}, \mathbf{G})$ and $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{R})$.

The objective is to verify whether the proposed procedures can correctly identify that the adopted model should include fixed and random intercepts and slopes.

We set $\sigma^2 = 1$ and $\rho = 0.6$ and considered 80 different settings (see Table 2) based on combinations of different values for $n, \sigma_a^2, \sigma_b^2$ and ρ_{ab} . For each setting we simulated 1000 samples and in each case we fitted second degree polynomials to the individual profiles and to the rows of the sample covariance matrices. We considered a significance level of 5% for the selection of fixed coefficients because the average profile plot is already a useful tool for such purpose; given that tools for the selection of random coefficients are not so simple, we set the significance level at 1% to be more rigorous, especially for the smaller sample sizes. The percentages of correct decisions (identification of first degree polynomials for the fixed and random coefficients) are presented in Table 2.

Table 2. Percentage of correct decisions for analysis based on individual profiles (ind) and rows of the covariance matrices (cov).

Settings				A		B		C		Settings				A		B		C	
n	σ_a^2	σ_b^2	ρ_{ab}	ind	cov	ind	cov	ind	cov	n	σ_a^2	σ_b^2	ρ_{ab}	ind	cov	ind	cov	ind	cov
5	4	2	0.75	74	56	78	72	82	81	25	1	1	0.75	94	92	80	98	80	98
5	4	2	0.25	72	60	78	74	81	81	25	1	1	0.25	94	96	80	98	81	98
5	2	2	0.75	71	61	78	75	80	81	25	1	0.5	0.75	85	68	80	94	80	97
5	2	2	0.25	70	60	78	76	80	84	25	1	0.5	0.25	80	74	80	95	79	98
5	2	1	0.75	51	41	67	56	76	73	25	0.5	1	0.75	95	96	81	98	80	98
5	2	1	0.25	48	44	65	58	75	69	25	0.5	1	0.25	94	96	80	98	82	98
5	1	2	0.75	72	63	78	77	80	84	25	0.5	0.5	0.75	85	75	80	95	82	97
5	1	2	0.25	72	63	77	78	82	85	25	0.5	0.5	0.25	83	75	81	96	80	98
5	1	1	0.75	47	41	66	59	75	71	50	4	2	0.75	95	100	77	99	76	98
5	1	1	0.25	47	42	65	61	74	72	50	4	2	0.25	95	100	76	99	77	98
5	1	0.5	0.75	24	21	48	39	59	53	50	2	2	0.75	94	100	77	99	79	98
5	1	0.5	0.25	23	23	47	41	62	55	50	2	2	0.25	95	100	76	99	77	98
5	0.5	1	0.75	49	41	65	61	72	72	50	2	1	0.75	95	99	76	99	76	99
5	0.5	1	0.25	46	45	65	62	73	75	50	2	1	0.25	95	100	76	99	78	99
5	0.5	0.5	0.75	23	23	48	42	61	57	50	1	2	0.75	94	100	77	99	76	98
5	0.5	0.5	0.25	21	24	46	43	59	55	50	1	2	0.25	94	100	76	99	77	98
10	4	2	0.75	93	82	83	93	83	96	50	1	1	0.75	94	100	76	99	78	99
10	4	2	0.25	93	85	83	93	84	95	50	1	1	0.25	96	100	76	99	74	99
10	2	2	0.75	92	86	84	94	86	96	50	1	0.5	0.75	92	92	77	99	77	99
10	2	2	0.25	92	88	83	94	82	96	50	1	0.5	0.25	94	93	77	99	76	99
10	2	1	0.75	80	62	81	83	82	92	50	0.5	1	0.75	95	100	76	99	77	99
10	2	1	0.25	80	69	81	84	82	90	50	0.5	1	0.25	94	100	76	99	76	98
10	1	2	0.75	93	88	83	94	84	95	50	0.5	0.5	0.75	94	91	77	100	79	99
10	1	2	0.25	92	88	84	94	85	96	50	0.5	0.5	0.25	94	94	77	99	78	98
10	1	1	0.75	81	65	82	85	81	92	100	4	2	0.75	95	100	85	100	87	99
10	1	1	0.25	82	71	82	86	82	92	100	4	2	0.25	96	100	84	100	85	99
10	1	0.5	0.75	51	38	75	67	80	81	100	2	2	0.75	95	100	84	100	86	99
10	1	0.5	0.25	55	45	74	68	83	81	100	2	2	0.25	95	100	85	100	84	99
10	0.5	1	0.75	81	69	82	87	80	93	100	2	1	0.75	95	100	85	100	83	100
10	0.5	1	0.25	80	71	82	87	82	92	100	2	1	0.25	95	100	85	100	86	99
10	0.5	0.5	0.75	53	43	75	69	81	85	100	1	2	0.75	95	100	85	99	84	98
10	0.5	0.5	0.25	50	46	75	69	79	84	100	1	2	0.25	95	100	85	99	84	99
25	4	2	0.75	95	99	80	98	78	98	100	1	1	0.75	95	100	85	100	85	100
25	4	2	0.25	94	99	81	98	82	98	100	1	1	0.25	95	100	85	100	84	99
25	2	2	0.75	95	99	80	98	80	98	100	1	0.5	0.75	95	99	85	100	84	100
25	2	2	0.25	95	99	80	98	78	98	100	1	0.5	0.25	93	99	84	100	85	100
25	2	1	0.75	93	92	80	98	80	98	100	0.5	1	0.75	94	100	84	100	86	99
25	2	1	0.25	94	95	80	98	79	98	100	0.5	1	0.25	96	100	85	100	84	100
25	1	2	0.75	93	99	80	98	79	97	100	0.5	0.5	0.75	95	99	84	100	83	100
25	1	2	0.25	95	100	79	98	80	97	100	0.5	0.5	0.25	95	100	85	100	85	99

The results in the columns corresponding to Assumption A in Table 2 suggest that it is more difficult to identify the correct model when σ_a^2 and σ_b^2 are smaller than σ^2 . In such cases, $\mathbb{V}(\hat{\beta}_i^*) = \sigma^2(\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1} + \mathbf{G}^*$ is dominated by $\sigma^2(\mathbf{X}_i^{*\top} \mathbf{X}_i^*)^{-1}$, masking the extra variability introduced by the presence of random coefficients. The high percentage (above 80%) of correct identification of the first degree polynomial for both fixed and random coefficients when an inappropriate model (a second degree polynomial) is adopted highlights the efficiency of the proposed procedure even for small sample sizes or misspecified underlying distributions, as suggested in the columns corresponding to the specifications B and C.

In the second study, we evaluated the performance of the proposed procedure in situations with missing observations. We considered similar settings as in the first study with three different values for ρ , namely, 0.4, 0.6 and 0.8. We assumed that each unit had a 50% probability of missing one observation; the position of the missing observation was chosen at random among the five available. In this study we did not consider the procedure based on the sample covariance matrix. The results, presented in Tables S1, S2 and S3 in the Supplementary Material and suggest that

- (i) for $n=5$ and $\sigma_a^2, \sigma_b^2 > \sigma^2$, the procedure leads to correct decisions in 50% of the cases when the data were generated according to specification A and in 50% to 60% of the cases for specifications B or C.
- (ii) for $n=5$ and $\sigma_a^2, \sigma_b^2 \leq \sigma^2$, the rate of correct decisions for specification A may be as low as 10%, while for specifications B and C, the percentage of correct decisions is of the order of 40% to 70%. This apparently unexpected result is possibly attributed to the fact that serial correlation may be confounded with random coefficients as suggested by Jones [7]. The extra variability imposed on the generated data by assuming serial correlation in specifications B and C is detected by the proposed procedure.
- (iii) for $n \geq 25$, the rate of correct decisions is greater than 70% for all specifications.

5. Analyses of real examples

The data in Table 3 (and in Table S4 in the Supplementary Materials) were extracted from a study conducted at the Heart Institute of the University of São Paulo (Incor), Brazil and is related to the growth of the systolic aorta diameter (measured echocardiographically) per unit weight of 29 pre-term neonates (PN) classified as adequate for gestational age (AGA) and of 32 PN classified as small for gestational age (SGA) according to their weight at birth. The PN were observed at irregularly spaced intervals (weeks) from birth to week 39 after presumed conception and one of the objectives was to estimate the corresponding growth curves. Details on the study may be obtained in [1].

The average and individual response profiles as well as a *loess* smoothed mean profile are displayed in Figures 1 and 2. The choice of a specific polynomial to represent the data is somewhat hampered because of the irregular pattern of the profiles, although quadratic polynomial models with random intercept, linear and quadratic terms may be a reasonable guess. We started by fitting a quadratic polynomial to each individual profile, excluding subjects with three or less observations: 6 subjects (ID 10, 12, 13, 15, 23 and 25 in Table 3) in the AGA group and 9 subjects (ID 34, 38, 43, 44, 45, 46, 49, 51 and 60 in Table S4 in the Supplementary Materials) in the SGA group. To reduce possible problems related to

Table 3. Aorta diameter per unit weight (mm/kg) of AGA PN.

Subject ID	Week post-conception														
	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	7.33	9.39	10.31		8.72	8.16	6.70	6.12		5.56		5.06			4.91
2					6.11	6.06	6.21		5.45	5.22	4.92				
3					5.75	6.26	5.78	5.07	4.88	4.58					
4				9.72	9.22		9.47		7.25			6.12	5.36	4.78	4.48
5							6.38	5.79	5.18	4.74					
6						5.44		4.94	4.57	4.32					
7			8.28	8.51	8.64		7.90		6.22		5.47		4.23		
8			7.67	8.60	7.90			6.57		5.74					
9									5.88	6.13		5.44			4.13
10								7.01	6.52						
11					5.20			4.76	4.22	4.07		3.69			
12				7.80	9.86										
13				6.84	7.73										
14							7.31	6.22		6.16	5.49	5.11		4.27	
15								5.56	5.39						
16						5.14	5.69	5.16	4.68	5.03	4.97				
17						6.09	6.61	6.12		4.70					
18					5.92	5.77	5.61		5.03						
19					4.82	6.30		6.89	6.17		5.38			4.46	
20							5.80		6.39	5.81	5.52		4.86		
21					7.17	6.94			5.74		4.82				
22			8.33	8.84	8.59	7.90		6.00							
23								4.84	4.76		4.40				
24					6.50	6.03	5.47		4.97						
25					6.17		5.90		5.51						
26					7.06	6.49	5.77		5.34		4.29				
27								5.78	5.76	5.21	5.03		4.58		
28							6.34	7.17		7.33			5.93		5.32
29					6.23	6.12		6.24			6.03		5.28		

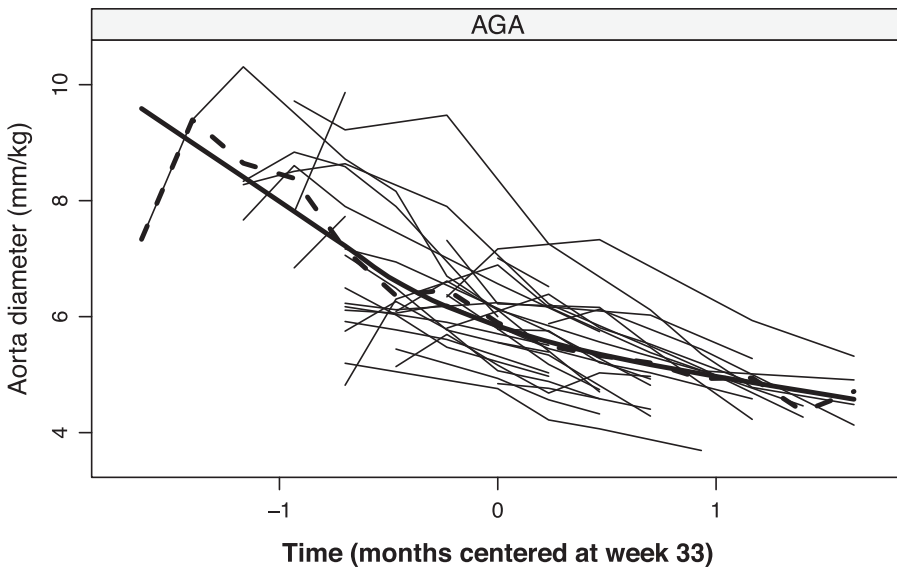


Figure 1. Profile plots for the aorta diameter per unit weight (mm/kg) of AGA PN (dashed line: mean profile; bold line: loess).

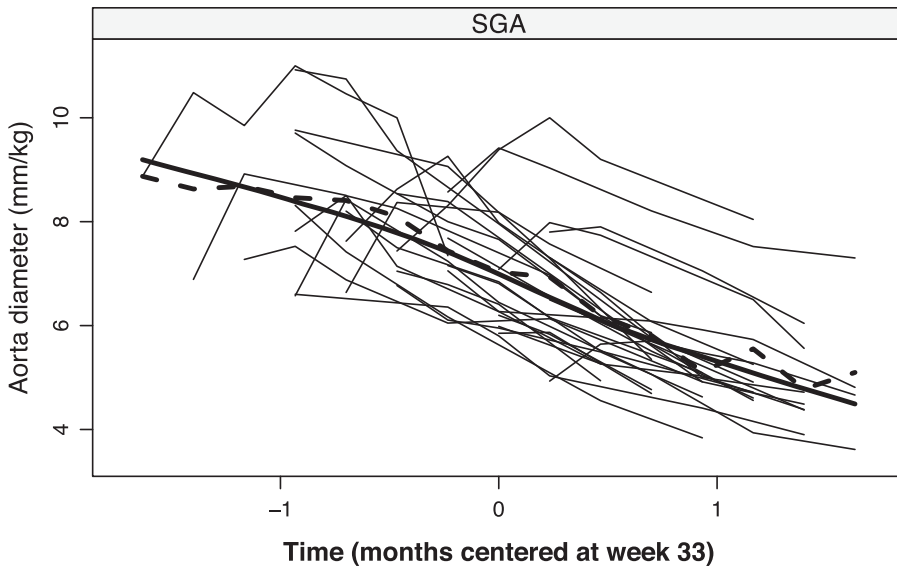


Figure 2. Profile plots for the aorta diameter per unit weight (mm/kg) of SGA PN (dashed line: mean profile; bold line: loess).

multicollinearity, we rescaled the time variable by taking $t_{ij}^* = (t_{ij} - 33)/4.29$ so that the time origin is week 33 and the time unit is month instead of week (note that $4.29 = 30 \text{ days} / 7 \text{ days per week}$). For the selection of the fixed coefficients we adopted a 5% significance level; for the construction of the reference Bonferroni-corrected confidence intervals used to identify random coefficients we adopted significance levels of 5% and also of 1%, to be more rigorous.

The parameter estimates and t -values for the quadratic polynomials fitted to the subjects in the AGA group are displayed in Table 4. Those corresponding to the subjects in the SGA group are presented in Table S5 in the Supplementary Materials.

Since only the means of the individual intercepts and slopes for the subjects in the AGA group are significant [$(\alpha = 1.67\% (= 0.05/3))$], these are the two terms that should be considered as candidates for fixed coefficients in the model. Adopting an overall significance level $\alpha = 5\%$, the Bonferroni-corrected individual significance level for the $69 = (23 \times 3)$ comparisons is $\alpha^* = 0.0003623 (= 0.05/69)$ so that the corresponding reference interval is $(-3.6, 3.6)$ derived from the t distribution with $\nu = 54 (= 123 - 69)$ degrees of freedom. From Table 4, we note that 52% ($= 12/23$) of estimated individual intercepts, 13% ($= 3/23$) of the estimated individual slopes and no estimated quadratic coefficients fall outside of Bonferroni-corrected reference interval, suggesting that the intercepts and slopes are the only two candidates for random coefficients. To be more rigorous, we also considered an overall significance level $\alpha = 1\%$ for which the corresponding Bonferroni-corrected reference interval is $(-4.1, 4.1)$, obtaining the same conclusions. With a confirmatory spirit, we fitted models with both fixed and random intercepts, slopes and quadratic coefficients to the data and considered tests of hypotheses to verify whether the quadratic terms could be dropped from the model. A standard test for the fixed quadratic coefficient yielded $p = 0.30$; for the variance component corresponding to the quadratic term, we considered

Table 4. Estimates of the individual intercepts, slopes and quadratic coefficients and corresponding *t*-values for the subjects in the AGA group (aorta diameter example).

Subject ID	Intercept		Slope		Quadratic	
	Estim	<i>t</i> -value	Estim	<i>t</i> -value	Estim	<i>t</i> -value
1	6.90	11.82	-1.50	-6.56	-0.06	0.32
2	5.84	-0.86	-0.92	0.14	-0.69	0.09
3	5.37	0.19	-1.53	-0.04	-0.94	0.07
4	7.89	11.34	-2.29	-5.68	-0.05	0.51
5	5.76	-1.79	-2.53	0.62	0.70	0.81
6	4.90	-5.16	-1.22	0.70	-0.12	0.62
7	7.21	5.36	-1.94	-5.51	-0.66	-1.00
8	6.88	0.94	-2.25	-0.64	-1.08	-0.45
9	5.91	-9.98	0.55	2.50	-1.01	-0.12
11	4.58	-2.90	-0.96	0.22	-0.04	0.61
14	6.68	5.49	-1.63	0.58	-0.05	0.59
16	5.14	-0.53	-0.44	1.61	0.14	0.75
17	6.26	-2.04	-1.59	0.23	-3.86	-0.74
18	5.35	-3.32	-1.21	1.43	-0.57	0.45
19	6.32	5.09	0.45	3.29	-1.39	-2.28
20	6.04	1.60	0.37	2.61	-1.24	-0.50
21	6.19	-4.51	-1.71	-1.30	-0.38	0.39
22	5.96	-5.03	-6.07	-0.85	-3.39	-1.44
24	5.19	-3.71	-1.20	1.44	1.01	1.03
26	5.54	-2.55	-1.88	-1.70	0.27	0.88
27	5.86	-4.37	-1.23	1.67	0.10	0.72
28	6.91	5.61	1.05	3.29	-1.32	-1.19
29	6.29	-0.73	-0.16	1.94	-0.56	-0.12
Estimate	6.04		-1.30		-0.66	
Std. error	0.09		0.23		0.39	
<i>p</i> -value	< 0.001		< 0.001		0.094	

the test proposed by Stram and Lee [17], obtaining $p = 0.94$. Both results suggest that the first degree polynomials selected for both the fixed and the random components of the model for the AGA group are adequate.

A similar analysis was performed for the SGA group. The parameter estimates and *t*-values for the quadratic polynomials fitted to the subjects in the SGA group are displayed in Table S5 in the Supplementary Materials and suggest that fixed and random intercepts, linear and quadratic coefficients should be included in the selected random coefficient regression model.

We fitted a joint model for both the AGA and SGA groups incorporating the suggestions from the exploratory analysis. The corresponding results are displayed in Table 5 under the label ‘Model 1’. This model suggests that the aorta diameter growth patterns differ according to whether the PN are classified as having weight adequate or not to the gestational age at birth.

We also compared the results obtained by fitting the model suggested by the proposed procedure with those of 21 homoskedastic conditionally independent competitors obtained from a comprehensive model by omitting or combining some terms. The terms included in each model are indicated in Table S6 of the Supplementary Materials. The comprehensive model is

$$\begin{aligned}
 y_{ijk} &= \alpha_i + \beta_i t_{ijk}^* + \gamma_i t_{ijk}^{*2} + a_{ij} + b_{ij} t_{ijk}^* + c_{ij} t_{ijk}^{*2} + e_{ijk}, \\
 i &= 1, 2, \quad j = 1, \dots, n_i, \quad k = 1, \dots, m_{ij},
 \end{aligned}
 \tag{9}$$

Table 5. Parameter estimates for competing models for the aorta diameter exemple (index 1: AGA and index 2: SGA).

Parameter	Model 1		Model 4		Model 8		Model 22	
	Estim	Std error	Estim	Std error	Estim	Std error	Estim	Std error
intercept (α_1)	6.04	0.15	6.04	0.17	6.09	0.18	6.04	0.15
intercept (α_2)	7.04	0.18	6.95	0.16	6.91	0.16	7.04	0.18
slope (β_1)	-1.23	0.12	-1.30	0.15	-1.20	0.18	-1.22	0.12
slope (β_2)	-1.36	0.2	-1.57	0.13	-1.63	0.13	-1.36	0.20
quadratic (γ_1)	-	-	-	-	-0.17	0.12	-	-
quadratic (γ_2)	-0.33	0.12	-	-	-	-	-0.33	0.12
$V(a) = \sigma_a^2$	-	-	0.712	-	0.779	-	-	-
$V(b) = \sigma_b^2$	-	-	0.287	-	0.563	-	-	-
$V(c) = \sigma_c^2$	-	-	-	-	0.085	-	-	-
$Cov(a, b) = \sigma_{ab}$	-	-	-0.074	-	0.062	-	-	-
$Cov(a, c) = \sigma_{ac}$	-	-	-	-	-0.115	-	-	-
$Cov(b, c) = \sigma_{bc}$	-	-	-	-	-0.203	-	-	-
$V(a_1) = \sigma_{a_1}^2$	0.565	-	-	-	-	-	0.599	-
$V(a_2) = \sigma_{a_2}^2$	0.893	-	-	-	-	-	0.894	-
$V(b_1) = \sigma_{b_1}^2$	0.191	-	-	-	-	-	0.212	-
$V(b_2) = \sigma_{b_2}^2$	0.926	-	-	-	-	-	0.931	-
$V(c_1) = \sigma_{c_1}^2$	-	-	-	-	-	-	0.008	-
$V(c_2) = \sigma_{c_2}^2$	0.128	-	-	-	-	-	0.129	-
$Cov(a_1, b_1) = \sigma_{ab_{11}}$	-0.232	-	-	-	-	-	-0.208	-
$Cov(a_1, c_1) = \sigma_{ac_{11}}$	-	-	-	-	-	-	-0.045	-
$Cov(b_1, c_1) = \sigma_{bc_{11}}$	-	-	-	-	-	-	-0.01	-
$Cov(a_2, b_2) = \sigma_{ab_{22}}$	0.217	-	-	-	-	-	0.218	-
$Cov(a_2, c_2) = \sigma_{ac_{22}}$	-0.260	-	-	-	-	-	-0.261	-
$Cov(b_2, c_2) = \sigma_{bc_{22}}$	-0.277	-	-	-	-	-	-0.278	-
$V(e_{ijk}) = \sigma^2$	0.304	-	0.346	-	0.307	-	0.300	-
AIC	707.5	-	714.2	-	714.3	-	713.0	-
BIC	762.1	-	743.3	-	758.0	-	778.5	-

where y_{ijk} denotes the k th observation of the aorta diameter for the j th subject in the i th group ($i = 1$: AGA, $i = 2$: SGA), t_{ijk}^* denotes the time (in months centred at week 33) in which this observation occurred, α_i , β_i and γ_i denote the fixed intercept, slope and quadratic coefficients, a_{ij} , b_{ij} and c_{ij} , respectively represent the corresponding random coefficients associated to the j th subject in the i th group and e_{ijk} denotes a random error term. We assume that the covariance matrix for the random terms is unstructured with variances denoted by $\sigma_{a_i}^2$, $\sigma_{b_i}^2$ and $\sigma_{c_i}^2$ and covariances, by σ_{abi} , σ_{aci} and σ_{bci} . The variance of the random error term is denoted by σ^2 . Results for 3 other models selected among the 21 considered for comparison are also presented in Table 5. The proposed model has a better fit than all the alternative ones according to the AIC criterion but not according to the BIC criterion. A residual analysis along the lines suggested in [16] is recommended to shed further light on the model selection.

The data analysed in the second example were extracted from a study analysed by Singer and Andrade [15] and are presented for illustrative purposes. The relative mucociliary transportation speed of frog palates was observed 5, 10, 15, 20, 25, 30 and 35 minutes after they were immersed in a solution of hydrogen peroxide with a concentration of $16 \mu\text{M}$. The goal was to model the expected response as a function of time. The data are shown in Table 6.

Table 6. Relative mucociliary transportation speed.

Row	Time in minutes						
	5	10	15	20	25	30	35
1	0.88	0.73	0.61	0.58	0.61	0.48	0.52
2	0.87	0.69	0.50	0.42	0.38	0.38	0.44
3	1.43	0.98	0.70	0.54	0.43	0.41	0.43
4	1.30	1.00	0.67	0.68	0.53	0.67	0.62
5	0.92	0.86	0.88	0.85	0.79	0.66	0.82
6	1.21	1.01	0.75	0.79	0.69	0.70	0.82
7	0.68	0.86	0.86	0.59	0.57	0.61	0.66
8	0.75	0.74	0.96	0.87	1.00	0.90	1.14
9	0.97	0.84	0.71	0.83	0.67	0.73	0.72
10	1.08	1.12	0.86	1.02	0.85	0.90	0.89

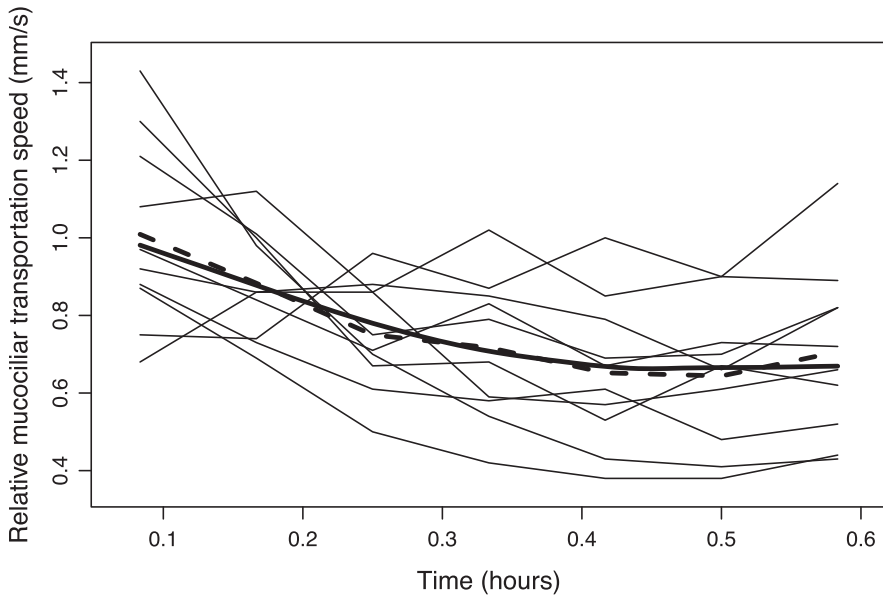


Figure 3. Profile plots for the relative mucociliary transportation speed data (dashed line: mean profile; bold line: loess).

To avoid multicollinearity problems, we rescaled the time variable, expressing it in terms of hours instead of minutes by taking $t_{ij}^* = t_{ij}/60$. The average and *loess* fitted profiles displayed in Figure 3 suggest that a quadratic polynomial may be adequate to represent the data. We fitted such a model to each individual response profile and considered a 5% significance level to select the candidates for both random and fixed coefficients. The results are displayed in Table 7.

All coefficients for the fixed coefficients were significant at the 5% level, confirming that a quadratic polynomial is adequate to represent the average profile. Furthermore, 40% (= 4/10) of the estimated individual intercepts, 30% (= 3/10) of the estimated individual slopes and 10% (= 1/10) of the estimated individual quadratic coefficients fall outside the corresponding Bonferroni-corrected reference interval $(-3.4, 3.4)$, obtained

Table 7. Estimates of the individual intercept, slope and quadratic coefficients and corresponding t -values (mucociliary speed data).

Unit	Intercept		Slope		Quadratic	
	Estim	t -value	Estim	t -value	Estim	t -value
1	1.00	-1.75	-1.84	0.68	1.75	-0.75
2	1.13	-0.53	-3.42	-1.42	3.82	1.14
3	1.87	6.19	-6.18	-5.09	6.43	3.50
4	1.66	4.24	-4.97	-3.48	5.62	2.77
5	0.98	-1.95	-0.67	2.24	0.50	-1.89
6	1.50	2.80	-3.84	-1.99	4.58	1.82
7	0.82	-3.42	-0.42	2.57	0.09	-2.26
8	0.70	-4.48	0.55	3.87	0.15	-2.20
9	1.07	-1.15	-1.56	1.05	1.70	-0.80
10	1.20	0.04	-1.17	1.58	1.10	-1.34
Estimate	1.19		-2.35		2.57	
Std. error	0.04		0.25		0.37	
p -value	< 0.001		< 0.001		< 0.001	

from the t distribution with 40 ($= 70 - 30$) degrees of freedom, suggesting that the intercept, slope and quadratic coefficients are also candidates for random coefficients. The same conclusion is reached if we use a 1% significance level for the decision. The corresponding Bonferroni-corrected reference interval is $(-3.9, 3.9)$.

The sample covariance matrix for the data in Table 6 is

$$S = \begin{bmatrix} 0.059 & 0.024 & -0.009 & -0.001 & -0.016 & -0.007 & -0.017 \\ 0.024 & 0.020 & 0.005 & 0.012 & 0.003 & 0.010 & 0.004 \\ -0.009 & 0.005 & 0.020 & 0.019 & 0.022 & 0.020 & 0.026 \\ -0.001 & 0.012 & 0.019 & 0.034 & 0.031 & 0.031 & 0.034 \\ -0.016 & 0.003 & 0.022 & 0.031 & 0.036 & 0.031 & 0.040 \\ -0.007 & 0.009 & 0.020 & 0.031 & 0.031 & 0.033 & 0.037 \\ -0.017 & 0.004 & 0.026 & 0.034 & 0.040 & 0.037 & 0.049 \end{bmatrix}. \quad (10)$$

The profile plots for the rows of (10), displayed in Figure 4, suggest that they may be represented by quadratic polynomials.

In Table 8 we present estimates and corresponding t -values for the coefficients of second degree polynomials fitted to the rows of (10) and observe that 57% of the estimated individual intercepts, 57% of the individual slopes and 14% of the estimated individual

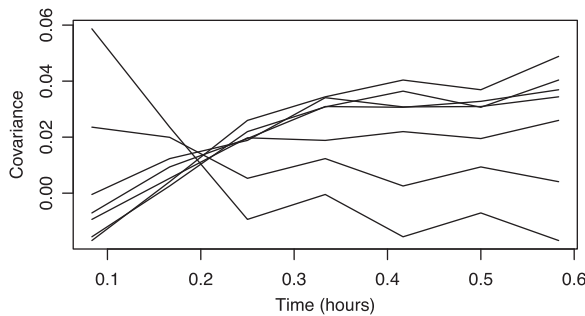


Figure 4. Profile plots of the rows of the sample covariance matrix (mucociliary speed data).

Table 8. Estimates of the intercept, slope and quadratic coefficients of the second degree polynomials fitted to the rows of the sample covariance matrix and corresponding *t*-values (mucociliary speed data).

Subject	Intercept		Slope		Quadratic	
	Estim	<i>t</i> -value	Estim	<i>t</i> -value	Estim	<i>t</i> -value
1	0.09	13.08	-0.008	-11.57	0.00014	8.30
2	0.04	5.05	-0.002	-4.35	0.00003	2.86
3	-0.03	-2.76	0.003	2.03	-0.00006	-1.63
4	-0.02	-1.86	0.003	2.14	-0.00006	-1.59
5	-0.04	-5.11	0.005	4.32	-0.00008	-3.04
6	-0.02	-2.96	0.004	2.69	-0.00006	-1.82
7	-0.04	-5.43	0.005	4.74	-0.00008	-3.11

quadratic coefficients fall outside the corresponding Bonferroni-corrected reference interval (-3.3, 3.3). This also suggests that the intercept, slope and quadratic coefficients are candidates for random coefficients. A standard test for the hypotheses of a null quadratic coefficient yielded $p < 0.003$; for the variance component corresponding to the quadratic term, we considered the test proposed by Stram and Lee [17], obtaining $p = 0.008$.

The estimated within-unit covariance matrix obtained from fitting the suggested model is given in Equation (11) and is quite close to the sample covariance matrix (10). The average relative absolute difference between their non-redundant elements is 1.6%.

$$\hat{V}(y_i) = \begin{bmatrix} 0.051 & 0.022 & 0.005 & -0.007 & -0.014 & -0.014 & -0.010 \\ 0.022 & 0.021 & 0.010 & 0.007 & 0.005 & 0.005 & 0.007 \\ 0.005 & 0.010 & 0.019 & 0.017 & 0.019 & 0.021 & 0.021 \\ -0.007 & 0.007 & 0.017 & 0.030 & 0.030 & 0.031 & 0.030 \\ -0.014 & 0.005 & 0.019 & 0.030 & 0.041 & 0.038 & 0.036 \\ -0.014 & 0.005 & 0.021 & 0.031 & 0.038 & 0.046 & 0.039 \\ -0.010 & 0.007 & 0.021 & 0.030 & 0.036 & 0.039 & 0.042 \end{bmatrix}. \quad (11)$$

We compared the proposed model to alternative ones as in the first example. The results, displayed in Table 9 also suggest that the proposed model has a better fit than the competitors.

Table 9. Parameter estimates for competing models (mucociliary speed example).

Parameter	Model 1		Model 2		Model 3		Model 4	
	Estim	Std error	Estim	Std error	Estim	Std error	Estim	Std error
intercept (α)	0.98	0.06	1.19	0.12	0.82	0.04	1.19	0.07
slope (β)	-0.64	0.21	-2.35	0.69	-0.03	0.11	-2.35	0.37
quadratic (γ)	-	-	2.57	0.76	-	-	2.57	0.46
$V(a_i) = \sigma_a^2$	0.03	-	0.13	-	0.26	-	0.03	-
$V(b_i) = \sigma_b^2$	0.36	-	4.14	-	9.04	-	0.39	-
$V(c_i) = \sigma_c^2$	-	-	4.42	-	10.38	-	-	-
$Cov(a_i, b_i) = \sigma_{ab}$	-0.080	-	-0.69	-	-1.49	-	-0.09	-
$Cov(a_i, c_i) = \sigma_{ac}$	-	-	0.72	-	1.6	-	-	-
$Cov(b_i, c_i) = \sigma_{bc}$	-	-	-4.27	-	-9.68	-	-	-
$V(e_{ij}) = \sigma^2$	0.01	-	0.01	-	0.01	-	0.01	-
AIC	-	-47.6	-	-86.7	-	-79.2	-	-70.6
BIC	-	-34.3	-	-64.7	-	-59.2	-	-55.2

To understand how profile plots like those in Figure 3 may be used in the process of selecting the random coefficients, we simulated data with the same fixed coefficients as those estimated under the model adopted for the mucociliary speed data with the following structure for the random coefficients: (i) only intercept, (ii) intercept and slope and (iii) intercept, slope and quadratic coefficient. The plots are presented in Figures S1, S2 and S3 in the Supplementary Materials and exhibit patterns corresponding to polynomials of the same degrees with which the data were generated. This indicates that examination of such plots may suggest the degree of the polynomial with which to start the analysis.

6. Discussion

In many problems where linear or nonlinear mixed models constitute the appropriate alternative to represent longitudinal data, the underlying information required for their specification is not available. In such cases, linear mixed models may be employed as reasonable approximations, provided the range for which we expect to conduct inference is limited. Such models are quite flexible and extensively studied in the statistical literature. Furthermore, stable and efficient software is available for their computational implementation. The associated flexibility, however, poses some difficulties for the practitioner, specially with respect to the choice of the fixed and random terms to be included in the model.

Selection of the appropriate model may depend on subject matter specific information which is lacking in many practical problems and one must rely on statistical tools to choose a reasonable one. The choice of the fixed coefficients can be handled quite simply by an examination of profile plots and well established hypothesis tests. Selection of the appropriate covariance structure, on the other hand, is not so straightforward. Likelihood ratio tests are not recommended to compare models with different fixed and random coefficients in view of REML estimation of the covariance parameters. Furthermore, AIC and BIC criteria are also subject to debate as mentioned in [9], so that the available selection tools should be used in a complementary manner.

We propose some simple exploratory tools that may be used along with other methods to identify potential candidates. In particular, they include fitting simple linear regression models to the individual response profiles as well as to the rows of the sample covariance matrix when it is available (for balanced data) and examining the distribution of the estimated coefficients.

Even for moderate sample sizes, the proposed tools may be useful, as shown by a limited simulation study. For samples of 25 units observed at 5 points in time, for example, the correct model was identified in around 80% or more cases, both with the analysis of the individual response profiles and of the rows of the sample covariance matrix. In two examples with data obtained from practical problems, the identified model fitted the data adequately. In particular, for the example with balanced data, the random terms included in the selected model were able to reproduce the sample covariance matrix up to reasonable differences.

In many instances, a model selected naively, either by experience with similar data or by the examination of profile plots may lead to the appropriate choice. This is clear from the plots in Figure 3. In cases like those depicted in Figures 1 and 2, however, the decision is not so straightforward. The exploratory tools we consider may not be a panacea for all

problems but may help in such situations and should be employed in the spirit of residual analysis to reassure that the chosen model is a reasonable one. Also, they are restricted to homoskedastic conditional independence linear mixed models; extension to other classes, like generalized linear mixed models is a challenging problem, given the ample range of functions that may be employed to represent the data. This seems to be a promising, although difficult research topic.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant 3304126/2015-2) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, grant 2013/21728-2), Brazil.

References

- [1] J.Y. Afune, *Echocardiographic evaluation of the evolution of preterm newborns from birth to term*, Ph.D. diss., (in Portuguese), Instituto da Criança, Universidade de São Paulo, Brazil, 2000.
- [2] E. Demidenko, *Mixed Models: Theory and Applications with R*, 2nd ed., Wiley, New York, 2013.
- [3] P. Diggle, P. Heagerty, K.-Y. Liang, and S Zeger, *Analysis of Longitudinal Data*, Oxford University Press, New York, 2002.
- [4] T. Fearn, *A two-stage model for growth curves which leads to Rao's covariance adjusted estimators*, *Biometrika* 64 (1977), pp. 141–152.
- [5] V. Giampaoli and J.M. Singer, *Likelihood ratio tests for variance components in linear mixed models*, *J. Statist. Plann. Inference* 139 (2009), pp. 1435–1448.
- [6] J.J. Grady and R.W. Helms, *Model selection techniques for the covariance matrix for incomplete longitudinal data*, *Stat. Med.* 14 (1995), pp. 1397–1416.
- [7] R.H. Jones, *Serial correlation or random subject effects*, *Comm. Statist.* 19 (1990), pp. 1105–1123.
- [8] N.M. Laird and J.H. Ware, *Random-effects models for longitudinal data*, *Biometrics* 38 (1982), pp. 963–974.
- [9] S. Müller, J.L. Scaely, and A.H. Welsh, *Model selection in linear mixed models*, *Statist. Sci.* 28 (2013), pp. 135–167.
- [10] J.G. Orelie and L.J. Edwards, *Fixed-effect variable selection in linear mixed models using statistics*, *Comput. Statist. Data Anal.* 52 (2008), pp. 1896–1907.
- [11] J.C. Pinheiro and D.M. Bates, *Mixed-effects Models in S and S-plus*, Springer-Verlag, New York, 2000.
- [12] W. Pu and X.-F. Niu, *Selecting mixed-effects models based on a generalized information criterion*, *J. Multivariate Anal.* 97 (2006), pp. 733–758.
- [13] M.N. Rao and C.R. Rao, *Linked cross-sectional study for determining norms and growth rates – a pilot survey of Indian school-going boys*, *Sankhya B* 28 (1966), pp. 237–258.
- [14] C.M. Rutter and R.M. Elashoff, *Analysis of longitudinal data: Random coefficient regression modelling*, *Stat. Med.* 13 (1994), pp. 1211–1231.
- [15] J.M. Singer and D.F. Andrade, *Analysis of longitudinal data*, in *Handbook of Statistics, Volume 18: Bio-Environmental and Public Health Statistics*, P.K. Sen and C.R. Rao, eds., North Holland, Amsterdam, 2000, pp. 115–160.
- [16] J.M. Singer, F.M.M. Rocha, and J.S. Nobre, *Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures*, *Int. Statist. Rev.* (2016), doi:10.1111/insr.12178, accepted for publication.
- [17] D.O. Stram and J.W. Lee, *Variance components testing in the longitudinal mixed effects model*, *Biometrics* 50 (1994), pp. 1171–1177.

- [18] E. Suyama, *Identificação de um modelo de efeitos aleatórios*, Ph.D. diss., (in Portuguese), Departamento de Estatística, Universidade de São Paulo, 1995.
- [19] G. Verbeke and G. Molenberghs, *Linear Mixed Models for Longitudinal data*, Springer, New York, 2000.
- [20] R.E Weiss and C.G. Lazaro, *Residual plots for repeated measures*, Stat. Med. 11 (1992), pp. 115–124.