1 Introduction

In the last decades, digital development has drastically changed the methodology of biomedical data analysis (Peek et al. [2014], Bellazzi et al. [2011]). On one hand, an extended availability of specific software has brought ease to the formerly tedious work of data entry for health records (Doukas et al. [2010]). Also, computers and servers today are capable of storing databases with a great number of features. Even if these variables are not relevant at the time of data collection, their values may help to explain odd or abnormal responses. (Koh et al. [2011])

Furthermore, developments in computational power and algorithm efficiency provide answers to research questions that decades ago would have been too expensive to solve from a financial, temporal and computational point of view. Nowadays, these questions can be researched and draw new conclusions at a considerably reduced expense (Hansen et al. [2014], Yoo et al. [2012]).

Given these issues, healthcare professionals can analyze follow-up data of a number of subjects with the same condition, and search for population and individual patterns (Cowie et al. [2017], Lau et al. [2012]) corresponding to a variable of interest (noted a response variable or outcome). Follow-up data are naturally longitudinal, and due to correlation between repeated measures (noted as measurement occasions), any statistical procedure that requires independent observations should be avoided to gain precision.

Longitudinal studies are often confused with time series. However, the main difference is that longitudinal studies have few measurement occasions compared to the number of subjects, whereas time series have few subjects with many observations per individual. This difference is crucial because the mathematical and statistical tools used for each discipline can be very different (Fitzmaurice et al. [2012]).

A frequent problem in longitudinal data analysis is that there is usually poor compliance regarding the study due to difficulties in measuring different subjects at exactly the same time or getting the subject to participate until the study is finished. This yields some missing values in the data, which is difficult to assess with verifiable assumptions (Newman [2003]).

For this reason, longitudinal data are usually unbalanced over time, meaning that not all responses are measured at the same time and not all subjects have the same number of measurements. Balanced longitudinal data are more common on experimental studies with strict protocols, and even in those cases, failure to completely collect the stipulated data is not a rare issue, whereas in observational studies it is almost a certain event.

A frequent characteristic of biomedical databases, is a high between-subject variability, since individuals usually respond differently in similar situations. More-

over, in longitudinal studies, the same subject is measured repeatedly over time and there can also be some within-subject variability.

Mixed models are widely accepted to analyze longitudinal data (Zhang and Davidian [2001]), since these models contemplate all individual and population response variability to attain a typical response for each individual, according to structures in the population response. The fitted values given by this model adapt to the heterogeneity in the data when well specified. This is possible because these models have individual random effects that can be fitted and interpreted. For example, a random intercept represents the individual baseline response (i.e., the indivual response at time zero) and a random slope denotes an individual growth rate of the outcome values over time.

Because of the heterogeneity of the data, it can be very difficult to answer some longitudinal research questions at a population level. That is why a natural approach is to find individuals whose response trajectory (the set of response variables of the same subject over time) is different from what is expected according to the population structure in the response variable. The difference in the response trajectory from these subjects can suggest the presence of a confounding variable not contemplated in the model, possibly leading to a new discovery (Suling and Pigeot [2012], Chawla and Davis [2013]).

Searching for these abnormal trajectories is essentially an unsupervised detection problem. The unsupervised nature is based on the fact that given a database, there is no data ascertaining the abnormality in the response trajectory. This lack of information poses enormous difficulty in the detection process, since the abnormal events are rare by definition and usually the details that differentiates the event are not always present.

This detection task has been extensively studied in the literature and can be found with several names: novelty detection, anomaly detection, outlier detection, fraud detection, depending on the area of application. Furthermore, Chandola et al. [2009] classifies three different types of anomalies: point anomalies, contextual anomalies and collective anomalies. The first category refers to single observations with values that are far away from the rest. The second category describes an observation that does not necessarily have an extreme value respect to others, but the value is far from the expected for a given context. Also, the last category applies to groups of observations that do not behave according to the majority of the dataset.

A common characteristic of large and high-dimensional data is that there is usually no prior knowledge regarding quantifiable relationships between variables. Thus, most outlier detection techniques described in the literature have minimal assumptions.

Some methods establish an outlier based on the density of near observations, by considering the distance between each observation and the rest and determining

the fraction of points exceeding a certain distance. If this proportion is high, the corresponding observation is far away from the rest of the data and can be labeled as an outlier. Conversely, if the number of observations in a certain neighbourhood is low, the same conclusion can be attained (Schubert et al. [2014]). Also, the decision can be based on the distance to the k-nearest neighbour (Ramaswamy et al. [2000]). However, in the corresponding research area, the Local Outlier Factor (LOF) is usually applied (Breunig et al. [2000]). Cluster-based methods function in a similar way. Objects that are not covered by a sufficient number of clusters or distant from all cluster centers are considered as outliers (Ester et al. [1996]).

Also, these approaches have univariate and multivariate options, the latter is mainly based on the Mahalanobis distance, which requires an estimation of the mean vector and covariance matrix (Rousseeuw and Van Zomeren [1990], Billor et al. [2000]). However, these estimations are dependent on distributional assumptions and are very susceptible to outliers and may fail to provide satisfactory results.

Furthermore, principal components analysis provide orthogonal linear combinations of variables that better explain the variance in the data, allowing to represent in few dimensions the majority of data points. However, a linear a combination of variables does not always have a direct interpretation, difficulting conclusions regarding specific variables.

These methods are effective to detect, based on a distance function, points or groups of points that are far away from the majority of the data. However, these observations can be considered normal depending on the interpretation of the data at the population level. Therefore, an addition of context can add a different perspective to a possible detection (Kriegel et al. [2008], Delannay et al. [2008]).

Another important issue is that these models require the same number of coordinates per observation, and therefore, complete data. Thus, if a certain subject has missing observations, the entire response trajectory (the set of responses for each subject) should be dismissed.

In some medical applications there is enough prior information on the variables of interest in order to propose a satisfactory statistical model for the data. A frequent approach is to assume that "normal" data follows this model, whereas data that differ from the predicted values exceeding a certain threshold is considered as abnormal. This approach is called semi-supervised learning, since the labelling of normal and abnormal data are based on the corresponding model, even if there is no certainty regarding abnormalities.

For fixed linear regression, many results use this approach. However, the major issue is in the estimation of parameters, since the presence of outliers can seriously influence these values, leading to masking and swamping effects. The masking effect occurs when a method fails to detect an outlier as such by affecting the fitted values. Conversely, under the swamping effect, an algorithm labels as

an outlier a normal observation, due to the change in the parameter values. Therefore, most outlier detection algorithms require robust parameter estimates, which are more time-consuming and usually require specific designs for a given application (Hardin and Rocke [2004], Leroy and Rousseeuw [1987]).

The boundaries used to classify extreme data can be based on a fixed distance between an expected value and the observed value, but can also rely on different dispersion measures of the residual values. Davies and Gather [1993] perform a detailed analysis of outlier identification for univariate data, assuming a gaussian distribution and based on two dispersion measures, the Standard Deviation and the Median Absolute Deviation (MAD). Sim et al. [2005] extend the work of Davies et al. to asymmetric distributions using use the Boxplot rule, based on the interquartile range (IQR).

Also, many papers have dealt with outlier detection in time series: Abraham and Box [1979], Fox [1972], Bianco et al. [2001], Roberts [2000], Lin et al. [2005], Tsay et al. [2000]. However, based on the aforementioned reasons, most of the corresponding methods are not applicable to a longitudinal setting.

Regarding mixed models, the literature usually focuses on the detection of subjects that influence the parameter estimates. However, an influential subject or observation does not necessarily mean that the corresponding response is abnormal.

Taking these issues into account, this work proposes using mixed effects models to identify abnormalities in a response trajectory. The proposal is based on the following advantages: first of all, high residual errors represent abrupt changes in the response values at a given time. This can be seen as in Chandola et al. [2009] as a contextual anomaly.

On the other hand, individuals with extreme random coefficients correspond to response trajectories morphologically similar to the population structure but with a distinct temporal evolution, depending on the interpretation of the coefficient. For example, an extreme random intercept denotes an abnormal baseline response and an extreme random slope equals to a different growth rate. According to Chandola et al. [2009] these are collective anomalies.

Moreover, these conceptually different notions of abnormal trajectories (contextual and collective) can be identified simultaneously using mixed models. This perspective is addressed in Zewotir and Galpin [2007], where the authors establish a constant threshold for extreme residual detection and identify extreme random effects using a standarization based on the estimated variance of each random effect.

Figure 1 allows to visualize the aforementioned anomalies. An anomaly will differ noticeably from a reference value, given by either the expected individual behaviour (extreme residual) or the expected population behaviour (extreme random intercept or slope). The response trajectories of this group of subjects have similar behaviour. The subject with an extreme residual has a trend over time consistent

with the population response except for a certain timepoint. The subject with an extreme random intercept has a response growth similar to the population, but starting from a higher value. Also, the subject with an extreme random slope shows a decrease in the response value, whereas the rest of the population is increasing their value.

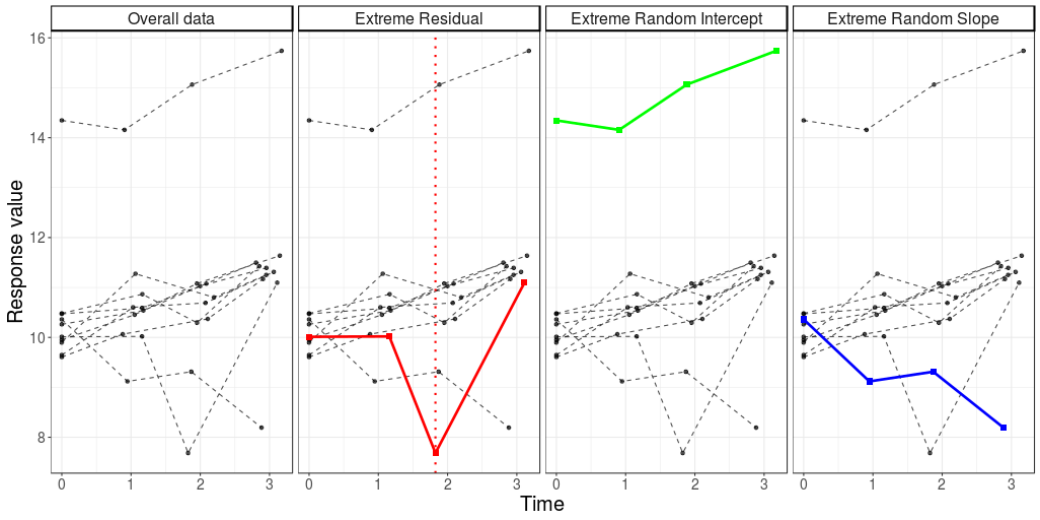


Figure 1: Examples of different anomalies.

The paper is organized as follows: Section 2.1 starts with some mathematical definitions regarding mixed effect models, followed by a description of the proposed algorithms in Section 2.2. Also, other methods introduced for the sake of comparison are described in Section 2.2.3. Details regarding the simulations are presented in Section 3, Section 3.1 presents the reference mixed model, Section 3.2 describes how abnormalities are included, Section 3.3 sets the parameter values for the simulations. Section 3.4 explains how missing data is introduced in a reference database and 3.5 and the proposed measures used to evaluate the performance of the algorithm are given in Section 3.5 outlines the measures used to evaluate the performance of the algorithm. The simulation results are presented in Section 3.6 and a comment regarding the influence of sample size is detailed in Section 3.8. The algorithm was tested in real databases described in Section 4 and the corresponding results are presented in Section 4.4. Finally, a discussion of the results is developed in Section 5.

2 Materials and methods

In the current section, the individuals considered in a study will be noted with index i ($1 \leq i \leq N$), each of which has n_i repeated measurements and the index j will describe each measurement occasion ($1 \leq j \leq n_i$), yielding a total of $N_T = \sum_{i=1}^{N} n_i$ observations. Also, we use $R^{n \times m}$ as the set of matrices with n rows and m columns. Another used notation is $\boldsymbol{I}_n$ denoting the $n \times n$ identity matrix. Matrices and vectors are noted in bold font.

The letter $\boldsymbol{Y}$ is used to refer to the response variable. For example, Y_{ij} represents the response variable for subject number i at measurement occasion j, whereas $\boldsymbol{Y}_i \in R^{n_i \times 1}$ is a vector containing all response measurements for subject number i, also called a response trajectory.

For simplicity, many equations will be expressed only in terms of index i, taking into account all repeated measures of a given subject, but assuming the equation as extensive to all N individuals.

2.1 Mixed Effect Models

A mixed effect model includes both fixed and random effects in the response variable. In matrix notation, fixed effects are usually defined with greek letter $\boldsymbol{\beta}$ as a $p \times 1$ vector with a corresponding design matrix $\boldsymbol{X}_i \in R^{n_i \times p}$, whereas random effects for subject number i is noted as $\boldsymbol{b}_i \in R^{q \times 1}$ with a design matrix given by $\boldsymbol{Z}_i \in R^{n_i \times q}$.

Using these definitions, linear mixed effects models have the following matrix formulation:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \times \boldsymbol{\beta} + \boldsymbol{Z}_i \times \boldsymbol{b_i} + \boldsymbol{\varepsilon}_i \tag{1}$$

where $\boldsymbol{\varepsilon}_i \in R^{n_i \times 1}$ represents the vector of measurement errors.

When applied to longitudinal databases, at least one of the columns of $\boldsymbol{X}_i$ represents a time-dependent variable that allows the identification of temporally ordered measures for the same subject. This condition usually also applies to $\boldsymbol{Z}_i$.

The model assumes independent multivariate normal distributions for $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$. Namely,

$$\left\{ \begin{array}{l} \boldsymbol{b}_i \sim N_q(\boldsymbol{0}; \boldsymbol{G}) \\ \boldsymbol{\varepsilon}_i \sim N_{n_i}(\boldsymbol{0}; \sigma^2 \boldsymbol{I}_{n_i}) \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} E[\boldsymbol{Y}_i] = \boldsymbol{X}_i \times \boldsymbol{\beta} \\ V[\boldsymbol{Y}_i] = \boldsymbol{Z}_i \times \boldsymbol{G} \times \boldsymbol{Z}_i' + \sigma^2 \boldsymbol{I}_{n_i} \end{array} \right\} \tag{2}$$

where $\boldsymbol{G}$ is a symmetric positive definite $q \times q$ matrix and $\boldsymbol{0}$ represents the null vector of the corresponding vector space. However, in practice, the model usually preforms well even when these assumptions are not totally met (see Verbeke and Lesaffre [1997]).

When these models are well specified, they are able to fit accordingly to between-subject variability via the random effects $\boldsymbol{b}_i$, without the use of one parameter per subject. This is because each of the individual vectors $\boldsymbol{b}_i$ is realization of the multivariate normal with zero mean and covariance matrix $\boldsymbol{G}$ common to all individuals which has an upper bound of $\frac{q(q+1)}{2}$ different values. Depending on the number of covariates associated with the random effects, the parameters involved can be considerably less than the N parameters that would be used by considering individual dummy variables (this number can increase if more than one parameter per individual is involved). However, the cost of parameter reduction is associated with a probabilistic structure imposed over the random effects that may not be valid.

Also, as it was mentioned before, longitudinal data usually exhibit a positive correlation between the repeated measures of a subject. The introduction of random effects in a model, induces a correlation between different measures of the same individual. Therefore, this issue is also contemplated by mixed models.

Furthermore, if the temporal variable considered is a continuous measure of time, it can adapt to missing values and unbalanced data, by adjusting a temporal structure that does not contemplate the specific time point with missing data.

Once the model is fitted, and $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}$ and $\widehat{\boldsymbol{G}}$ are estimated, a marginal or population response can be calculated as $\widehat{\boldsymbol{Y}_i^{\boldsymbol{0}}} = \boldsymbol{X}_i \times \widehat{\boldsymbol{\beta}}$. Also, the individual random coefficients $\boldsymbol{b}_i$ can be fitted using the empirical best linear unbiased predictor (Empirical BLUP) given by:

$$\begin{aligned} \widehat{\boldsymbol{b}}_i &= \widehat{\boldsymbol{G}}\boldsymbol{Z}_i'\widehat{\boldsymbol{H}}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}) \\ \widehat{\boldsymbol{H}}_i &= \boldsymbol{Z}_i\widehat{\boldsymbol{G}}\boldsymbol{Z}_i' + \widehat{\sigma}^2\boldsymbol{I}_{n_i} \end{aligned} \tag{3}$$

After obtaining $\widehat{\boldsymbol{b}}_i$, a fitted individual response trajectory for subject i is given by

$$\widehat{\boldsymbol{Y}}_i = \boldsymbol{X}_i \times \widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i \times \widehat{\boldsymbol{b}}_i. \tag{4}$$

This yields a residual vector denoted $\boldsymbol{r}_i = \boldsymbol{Y}_i - \widehat{\boldsymbol{Y}}_i \in R^{n_i \times 1}$. If some value r_{ij} ($1 \leq j \leq n_i$) is considered extreme (far away from zero), it means that the response for subject number i at measurement occasion number j has a peak in the response trajectory. This event can be identified in order to examine the reasons for such abrupt change in the response.

Assuming a statistical model for the data, a regular observation is considered to follow a certain distribution F, whereas anomalies or outliers are supposed to follow a different distribution F_c and therefore, generate abnormal data in an outlying region (Davies and Gather [1993]). A possible structure for observations from distribution F_c (also called contaminants) consists of the addition of a constant to

values from distribution F. This model for the outliers is called a mean shift outlier model (MSOM).

In this case, a regular observation $Y_{i,j}$ is expected to follow a normal distribution with mean $(\boldsymbol{X}_i \times \boldsymbol{\beta})_j$ and standard deviation $(\boldsymbol{Z}_i \times \boldsymbol{G} \times \boldsymbol{Z}_i')_{j,j}$, whereas for an outlying observation $Y_{i,j}$, the MSOM assumes that for a constant δ, $Y_{i,j}$ is considered to be following a normal distribution with mean $(\boldsymbol{X}_i \times \boldsymbol{\beta})_j + \delta$ and standard deviation $(\boldsymbol{Z}_i \times \boldsymbol{G} \times \boldsymbol{Z}_i')_{j,j}$.

Given these definitions, Figure 2 revisits the data from Figure 1 identifying the empirical response $\boldsymbol{Y}_i$, the fitted individual response $\widehat{\boldsymbol{Y}}_i$ and the marginal or population response $\widehat{\boldsymbol{Y}_i^{\mathbf{0}}}$. The subject with an extreme residual shows similar values to the fitted response except for a certain timepoint. The difference in the subjects with extreme random effects appears mainly between the empirical response and the marginal response.

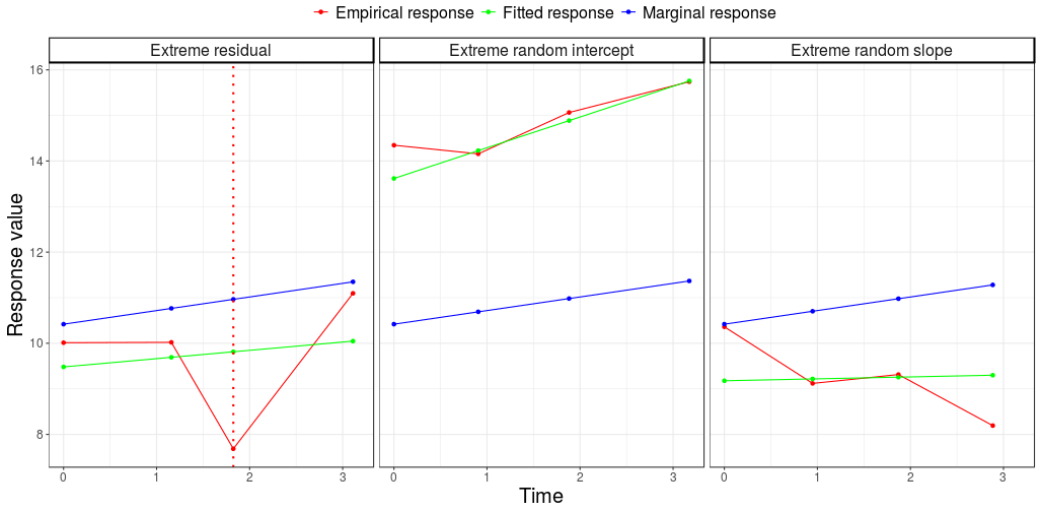


Figure 2: Depiction of how the different anomaly types are visualized and the corresponding behaviour of the fitted responses.

Especifying results from Zewotir and Galpin [2007], using the block structure of the matrices relative to longitudinal data, the residuals can be standardized calculating the corresponding covariance matrix:

$$\begin{gathered} \boldsymbol{r}_i = \boldsymbol{R}_i \times \boldsymbol{Y}_i \\ \boldsymbol{R}_i = \boldsymbol{H}_i^{-1} - \boldsymbol{H}_i^{-1} \times \boldsymbol{X}_i \times \left(\sum_{k=1}^{N} \boldsymbol{X}_k' \times \boldsymbol{H}_k^{-1} \times \boldsymbol{X}_k\right)^{-1} \times \boldsymbol{X}_i' \times \boldsymbol{H}_i^{-1} \\ \boldsymbol{r}_i \sim N_{n_i}\left(\mathbf{0}; \sigma^2 \cdot \boldsymbol{R}_i\right) \Rightarrow V(r_{ij}) = (\boldsymbol{R}_i)_{j,j} \end{gathered} \tag{5}$$

Therefore, the residual vectors can be standardized as follows:

$$t_{ij} = \frac{r_{ij}}{\widehat{\sigma} \cdot \sqrt{(\boldsymbol{R}_i)_{j,j}}} \tag{6}$$

Furthermore, the predicted residuals (excluding observation j of subject i in the estimation of σ) can also be analyzed:

$$\begin{aligned} t^*_{ij} &= \frac{r_{ij}}{\widehat{\sigma}_{(i,j)} \cdot \sqrt{(\boldsymbol{R}_i)_{j,j}}} \\ \widehat{\sigma}^2_{(i,j)} &= \widehat{\sigma}^2 \cdot \left(\frac{N_T - t^2_{ij}}{N_T - 1} \right) \end{aligned} \tag{7}$$

Verbeke and Molenberghs [2000] apply the calculations in Zewotir and Galpin [2007] to longitudinal data. According to this work the random effects follow a distribution with a specific covariance structure:

$$\widehat{\boldsymbol{b}_i} \sim N_q(\boldsymbol{0}; \boldsymbol{G} \times \boldsymbol{Z}'_i \times \boldsymbol{R}_i \times \boldsymbol{Z}_i \times \boldsymbol{G}) \tag{8}$$

Thus, $\widehat{\boldsymbol{b}_{ih}}$ (the estimated random effect number h for subject number i, where $1 \leq h \leq q$) can be standarized as follows:

$$\widehat{v_{ih}} = \frac{\widehat{b_{ih}}}{\sqrt{(\widehat{\boldsymbol{G}} \times \boldsymbol{Z}'_i \times \boldsymbol{R}_i \times \boldsymbol{Z}_i \times \widehat{\boldsymbol{G}})_{h,h}}} \tag{9}$$

These calculations were individualized according to each subject i. However, these definitions can be extended to the entire population, omitting the subscript i. For example, $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_N)$ is the $N_T \times 1$ vector including all trajectory responses, matrix $\boldsymbol{X} \in R^{N_T \times p}$ represents the individual design matrices $\boldsymbol{X}_i$ binded by rows, whereas the sparse matrix $\boldsymbol{Z} \in R^{N_T \times (Nq)}$ has a block structure with $\boldsymbol{Z}_i$ disposed diagonally. Also, the residual vectors $\boldsymbol{r}$ and $\boldsymbol{t}$ (standarized) have $N_T \times 1$ coordinates, whereas the estimated random effect matrices $\boldsymbol{B}$ and $\boldsymbol{V}$ (standarized) have N rows and q columns. $\boldsymbol{B}$ and $\boldsymbol{V}$ bind by rows the N vectors of q estimated individual random effects $\widehat{\boldsymbol{b}_i}$ and $\widehat{\boldsymbol{v}_i}$, respectively.

2.2 Algorithm

The algorithm that we propose next was implemented in R software, since this programming language provides an easy management of databases and handles models involving categorical data very well. The algorithm is divided according to each detection task, given by Algorithm 1 and Algorithm 2, shown in Sections 2.2.1 and 2.2.2, respectively.

Algorithm 1 Extreme Residual Detection Algorithm

procedure RESIDUAL DETECTION(*Data*, *Resp*, *Covs*, *Rands*, *ID*, T_P)
  $NetData \leftarrow$FilterMissingData$(Data, Resp, Covs, Rands, ID)$
  $Y \leftarrow NetData(Resp)$
  $NRows \leftarrow$ Length(Y)
  $X \leftarrow$DesignMatrix$(Covs, NetData)$
  $Z \leftarrow$DesignMatrix$(Rands, NetData, ID)$
  $Model \leftarrow$FitMixedModel(Y, X, Z)
  $Res \leftarrow$Residuals$(Model)$
  $L_P = Q1(Res) - T_P \cdot IQR(Res)$
  $U_P = Q3(Res) + T_P \cdot IQR(Res)$
  $List_{Res} \leftarrow \emptyset$
  $k \leftarrow 1.$
  loop:
  while $k \leq NRows$ **do**
    $Aux \leftarrow (ID(k), Time(k))$
    if $Res(k) < L_P$ **then**
      $List_{Res} \leftarrow$Append$(List_{Res}, Aux)$
    else if $Res(k) > U_P$ **then**
      $List_{Res} \leftarrow$Append$(List_{Res}, Aux)$
    $k \leftarrow k + 1.$
    go to *loop*.
  return $List_{Res}$

The algorithms' variables, inputs and outputs are noted in italic font. For example, *Res*, defined in Algorithm 1.

The algorithm input is given by the database *Data*, the name response variable *Resp*, a vector of strings *Covs* containing the predictive variable names, a vector of strings *Rands* containing the names of the variables associated to the random effects, the subject identification variable name *ID*, two threshold values T_P and T_R.

2.2.1 Extreme residuals

Based on these parameters, the algorithm only excludes the records that have missing data regarding the variables considered in the mixed model. Afterwards, it fits the model (following equation (1), named "FitMixedModel" in Algorithm 1) estimating the fixed coefficients $\widehat{\boldsymbol{\beta}}$, the residual standard error $\widehat{\sigma}$ and the random effects

covariance matrix $\widehat{\boldsymbol{G}}$. Based on these estimations it is possible to obtain the individual random effects $\widehat{\boldsymbol{b}}_i$, the fitted response trajectories $\widehat{\boldsymbol{Y}}_i$ (for observations without missing covariates) and the residuals $\boldsymbol{r}_i$.

The first detection task (Algorithm 1) constructs two boundaries based on the residual vector *Res*, using the corresponding first quartile ($Q1(Res)$), the third quartile ($Q3(Res)$), the respective interquartile range ($IQR(Res)$) and the threshold value T_P given as input.

Residual values under L_P or over U_P are considered extreme and the respective ID and measurement time is stored in a list.

Obviously, a higher value for T_P yields a more restrictive threshold for detecting outliers, whereas low values will tend to detect more data as abnormal.

2.2.2 Extreme Random Effects

In a similar way, naming q the number of random effects, Algorithm 2 takes the mixed model (noted as *Model*) fitted in Algorithm 1 as an input. Afterwards, calculates the matrix of estimated random effects $\boldsymbol{B} \in R^{N\times q}$ (one row per subject and one column per random effect). Each column of $\boldsymbol{B}$ is named as $\boldsymbol{B}_h \in R^{N\times 1}$, $1 \leq h \leq q$ in Algorithm 2. Also, $\boldsymbol{B}_h(i)$ refers to the estimated random effect number h for subject number i.

Using all these definitions, the second detection task (Algorithm 2), constructs q different ranges of "normal" values for the respective random effect. An estimated random effect $\boldsymbol{B}_h(i)$ ($1 \leq i \leq N$, $1 \leq h \leq q$) will be considered extreme if the value is not between L_h and U_h. Also, the respective ID, the name of the variable corresponding to random effect number h (*VarName*), the value $\boldsymbol{B}_h(i)$ (given by *Val* in Algorithm 2) is stored in a list.

The major difference between both tasks is that the former detects specific times of a response trajectory and the latter procedure detects entire response trajectories with extreme random effects values. Since the number of individuals (and thus, the number of response trajectories) is given by N and the total number of observations is $N_T = \sum_{i=1}^{N} n_i$ (much larger than N), any diagnostics measure involving the second detection task is more susceptible to errors than the first one.

2.2.3 Comparisons with other thresholding methods

As mentioned in Section 1, other dispersion measures can be used to define the range of "normal" data. One is the well known Standard Deviation (denoted *SD* in this work) and the more robust Median Absolute Deviation (MAD). Given a data

Algorithm 2 Extreme Random Effects Detection Algorithm

procedure RANDOM EFFECT DETECTION(*Data*, *Resp*, *Model*, *ID*, T_R)
  $Y \leftarrow Resp$
  $B \leftarrow$FitRandomEffects$(Model, Y)$
  $i \leftarrow 1$
  $h \leftarrow 1$
  $List_{Rand} \leftarrow \emptyset$
  loop1:
  while $h \leq q$ **do**
    $VarName \leftarrow Rands(h)$
    $B_h =$ExtractColumn(B, h)
    $L_h = Q1(B_h) - T_R \cdot IQR(B_h)$
    $U_h = Q3(B_h) + T_R \cdot IQR(B_h)$
    loop2:
    while $i \leq N$ **do**
      $Val \leftarrow B_h(i)$.
      if $B_h(i) < L_h$ **then**
        $Aux \leftarrow (ID(i), VarName, Val)$
        $List_{Rand} \leftarrow$Append$(List_{Rand}, Aux)$
      else if $B_h(i) > U_h$ **then**
        $Aux \leftarrow (ID(i), VarName, Val)$
        $List_{Rand} \leftarrow$Append$(List_{Rand}, Aux)$
      $i \leftarrow i + 1$.
      go to *loop2*.
    $h \leftarrow h + 1$.
    go to *loop1*.
  return $List_{Rand}$

vector $x = (x_1, \cdots, x_n)$, the MAD is calculated based on the median ($\tilde{x} =$Median(x)) as follows:

$$MAD(x) = k \cdot \text{Median}\{|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, \cdots, |x_n - \tilde{x}|\} \tag{10}$$

where $k = \frac{1}{\Phi^{-1}(0.75)}$ is a scale parameter. In this expression, $\Phi^{-1}(0.75)$ represents the third quartile of a standard normal distribution. This value of k ensures that, for normal data, $MAD(x)$ is an unbiased estimator of the standard deviation, obtained by the next equation, given the mean (Mean$(x) = \bar{x}$):

$$SD(x) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{11}$$

Thus, the algorithms given in the last section can be adapted to these new boundaries replacing:

Standard deviation:

- **Algorithm 1:**
  $L_P = Mean(Res) - T_P \cdot SD(Res)$
  $U_P = Mean(Res) + T_P \cdot SD(Res)$
- **Algorithm 2:**
  $L_h = Mean(B_h) - T_R \cdot SD(B_h)$
  $U_h = Mean(B_h) + T_R \cdot SD(B_h)$

where $Mean(Res)$ and $Mean(B_h)$ are the means of the residual vector Res and B_h (the column number h of fitted random effects matrix B) respectively. Also, $SD(Res)$ and $SD(B_h)$ are noted as the standard deviation of these vectors.

Median Absolute Deviation:

- **Algorithm 1:**
  $L_P = Med(Res) - T_P \cdot MAD(Res)$
  $U_P = Med(Res) + T_P \cdot MAD(Res)$
- **Algorithm 2:**
  $L_h = Med(B_h) - T_R \cdot MAD(B_h)$
  $U_h = Med(B_h) + T_R \cdot MAD(B_h)$

In an analogous way, $Med(Res)$ and $Med(B_h)$ are the medians of the residual vector $\boldsymbol{r}$ and columns $\boldsymbol{B}_h$ respectively, whereas $MAD(Res)$ and $MAD(B_h)$ are the Median Absolute Deviation of these vectors.

Remark: There is an issue worth mentioning. Both alternatives described above are symmetric respective to the measure of central tendency (mean and median respectively), whereas the first proposal, based on the boxplot rule, allows for skewness in the range of normal values, since the first and third quartile are not necessarily equidistant to the median.

Standarizations:

Also, the same thresholds can be applied replacing residual vector Res (mathematically defined as $\boldsymbol{r}$) by the standarized ($\boldsymbol{t}$) and predicted residuals ($\boldsymbol{t}^*$). In case of the estimated random effects, the matrix $\boldsymbol{B}$ can be replaced by $\boldsymbol{V}$.

Zewotir and Galpin [2007]:

In their paper, Zewotir and Galpin [2007] propose the following thresholds for both detection tasks, using the standarized residuals ($\boldsymbol{t}$) and random effects ($\boldsymbol{V}$):

- **Algorithm 1:**
  $L_P = -\sqrt{\frac{4 \cdot N_T}{N_T - p + 3}}$
  $U_P = \sqrt{\frac{4 \cdot N_T}{N_T - p + 3}}$
- **Algorithm 2:**
  $L_h = -t_{0.975,DF}$
  $U_h = t_{0.975,DF}$

Where $DF = N_T - \text{rank}[\boldsymbol{X}\ \boldsymbol{Z}] - 1$ and $\text{rank}[\boldsymbol{X}\ \boldsymbol{Z}]$ is the rank of matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ binded by columns. Also, it is worth mentioning that the work does not

focus on missing data and thus, N_T is assumed equal to the number of observations. Therefore, in the event of missing data, N_T should be replaced by the number of net observations N_O, and filtering the number of missing values of variables involved in the mixed model.

3 Simulated databases

To test the algorithm, we simulated data using a simple mixed model, with parameter values and variables that give rise to data with similar morphological characteristics to those found in clinical follow-up data.

3.1 Parameters and variables

With this goal in mind, a hypothetical study was considered with two groups (Control and Treatment, for example), with different fixed intercepts and time slopes, all included in a vector of fixed parameters noted as $\boldsymbol{\beta}$. Also, a random intercept and time slope model ($\boldsymbol{b}_i = (b_{i1}, b_{i2})$) was used, with zero mean and a defined covariance matrix $\boldsymbol{G}$. In addition, random measurements errors ε_{ij} were added.

$$\begin{aligned} &\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4) \\ &\boldsymbol{b}_i = (b_{i1}, b_{i2}) \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{G}) \\ &\boldsymbol{G}_{hl} = \mathrm{Cov}(b_{ih}, b_{il}) \quad (1 \leq h, l \leq 2) \\ &\varepsilon_{ij} \sim \mathcal{N}(0, \sigma) \end{aligned} \tag{12}$$

Considering that longitudinal data are usually unbalanced over time, a new approach was implemented for the construction of time measurements. Given that the number of measurements are usually not the same for all individuals, a random number of observation n_i ($J_{\min} \leq n_i \leq J_{\max}$) was assigned to subject number i.

To achieve mistimed measurements, given n_i, the measurement occasions were set as $t_{ij} = \sum_{k=1}^{j} \tau_{ik}$, where $1 \leq j \leq n_i$ and τ_{ik} are independent exponential variables of parameter $\lambda = 1$.

With all these parameters, each individual fixed effects design matrix $\boldsymbol{X}_i \in R^{n_i \times 4}$ is built setting the first column to a value of one, the second column is given by $\{t_{ij}\}_{1 \leq j \leq n_i}$. The third and fourth columns of $\boldsymbol{X}_i$ are identical to the first and second, respectively, if subject number i belongs to the Treatment group. Otherwise, the last two columns have a value of zero.

Since the treatment group is not included as a random effect, each matrix $\boldsymbol{Z}_i \in R^{n_i \times 2}$ is equal to the the first two columns of $\boldsymbol{X}_i$. Once all the matrices and coefficients are attained, a simulation for each response trajectory $\boldsymbol{Y}_i$ is calculated following equation (1).

3.2 Abnormalities

In order to include abnormalities to detect in the data, two different deviations $\boldsymbol{D}^P$ and $\boldsymbol{D}^R$ were added to the mixed model. First, $\boldsymbol{D}^P$ is added to the normal response trajectory $\boldsymbol{Y}_i$, where $\boldsymbol{D}^P$ has non-zero value with probability p_P. In order to be considered a peak, the deviation value should be relative to the variability of the measurement errors σ. With this in mind, another positive input value u_P is considered as a constant multiplying σ, used as the absolute value of the deviation. The variable $\boldsymbol{D}^P$ is built as follows:

$$\begin{aligned} &U_{ij}^1 \sim \mathscr{U}(0,1) \\ &U_{ij}^2 \sim \mathscr{U}(0,1) \\ &D_u = u_P \cdot \sigma \\ &D_{ij}^P = \begin{cases} D_u & \text{if } U_{ij}^1 \leq p_P, U_{ij}^2 > 0.5 \\ -D_u & \text{if } U_{ij}^1 \leq p_P, U_{ij}^2 \leq 0.5 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{13}$$

In this construction, U_{ij}^1 and U_{ij}^2 are random numbers of uniform distribution between 0 and 1.

It is worth noting that, to avoid introducing bias, the sign of $\boldsymbol{D}^P$ can be negative or positive in equal proportions.

To test the second detection task, a similar strategy is used only relative to the variability of each random effect. For example, knowing that

$$G_{11} = \text{Cov}(b_{i1}, b_{i1}) = \text{Var}(b_{i1}), \tag{14}$$

the random intercept variability is given by $\sqrt{G_{11}}$. To introduce an extreme random intercept, a variable $\boldsymbol{D}^{RI}$ is calculated as in (13), with different values and another probability of non-zero value p_{RI}. Also, a positive input parameter given by u_R determines the absolute value of deviation $\boldsymbol{D}^{RI}$ relative to the dispersion $\sqrt{G_{11}}$:

$$\begin{aligned} &D_u = u_R \cdot \sqrt{G_{11}} \\ &D_i^{RI} = \begin{cases} D_u & \text{if } U_i^1 \leq p_{RI}, U_i^2 > 0.5 \\ -D_u & \text{if } U_i^1 \leq p_{RI}, U_i^2 \leq 0.5 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{15}$$

Analogous to $\boldsymbol{D}^{RI}$, $\boldsymbol{D}^{RS}$ is calculated as in (15) using $\sqrt{G_{22}}$ instead of $\sqrt{G_{11}}$ and p_{RS} replacing p_{RI}. These values are included in a vector $\boldsymbol{D}^R = (\boldsymbol{D}^{RI}, \boldsymbol{D}^{RS})$ and the final simulated response for subject number i, given by $\widetilde{\boldsymbol{Y}}_i$ is calculated as:

$$\widetilde{\boldsymbol{Y}}_i = \boldsymbol{X}_i \times \boldsymbol{\beta} + \boldsymbol{Z}_i \times \boldsymbol{b}_i + \boldsymbol{D}_i^P + \boldsymbol{Z}_i \times \boldsymbol{D}_i^R + \boldsymbol{\varepsilon}_i \tag{16}$$

where $\boldsymbol{D}_i^P$ represents the vector consistent of the n_i values of $\boldsymbol{D}_{ij}^P$ calculated in Equation 13 ($1 \leq j \leq n_i$).

In order to appreciate the interaction effect between $\boldsymbol{D}^{RI}$ and $\boldsymbol{D}^{RS}$, a unique value $p_R = p_{RI} + p_{RS}$ is given as input of the simulation, and three combinations are considered:

- $p_{RI} = p_R,\ p_{RS} = 0$
- $p_{RI} = 0,\ p_{RS} = p_R$
- $p_{RI} = p_{RS} = \frac{p_R}{2}$

Also, whenever the notation u_R is used, u_{RI} and u_{RS} are considered to be equal ($u_R = u_{RI} = u_{RS}$). Otherwise, u_{RI} and u_{RS} can be specified with different values.

3.3 Simulation settings

Three settings were considered in order to evaluate three different detections: extreme residuals, extreme random intercepts and extreme random slopes. Specific parameter values were considered for each case. However, some parameters were fixed through all three experiments. Also, the parameters for the simulated longitudinal databases were selected with the aim of resembling biomedical follow-up data. These values are given in Table 1.

Table 1: Parameter values used for all simulations

β_1	1	β_2	0.1	β_3	-2	β_4	-0.2	$J_{\min}$	2
G_{11}	0.051	G_{12}	-0.001	G_{22}	0.051	σ	0.25	$J_{\max}$	4

Parameter values for extreme residual detection:

p_{RI} : 0, 0.025, 0.05	p_{RS} : 0, 0.025, 0.05	p_P : 0.05, 0.1, 0.2, 0.25
u_{RI} : 1, 2	u_{RS} : 1, 2	u_P : 3, 4
T_{RI} : 4	T_{RS} : 4	N : 100, 200, 300

Parameter values for extreme random intercept detection:

p_P : 0, 0.1, 0.2	p_{RS} : 0, 0.05, 0.1, 0.2	p_{RI} : 0.05, 0.1, 0.2
u_P : 1, 2	u_{RS} : 1, 2	u_{RI} : 3, 4
T_P : 4	T_{RS} : 4	N : 100, 200, 300

Parameter values for extreme random slope detection:

p_P : 0, 0.1, 0.2	p_{RI} : 0, 0.05, 0.1, 0.2	p_{RS} : 0.05, 0.1, 0.2
u_P : 1, 2	u_{RI} : 1, 2	u_{RS} : 3, 4
T_P : 4	T_{RI} : 4	N : 100, 200, 300

According to Kannan et al. [2015], for normal distributions, the boundaries L_P and U_P calculated in Section 2.2 with the IQR ($T_P^{IQR} = 1.5$) are similar to those obtained in Section 2.2.3 with the standard deviation and the median absolute deviation ($T_P^{SD} = T_P^{MAD} = 3$). Therefore, in each setting, the comparisons between these thresholding methods will be made with these values for T_P ($T_P^{IQR} = 1.5$ and $T_P^{MAD} = T_P^{SD} = 3$). The respective comparisons in other settings are established replacing T_P by T_{RI} (or T_{RS}), maintaining the same values for the other different detection tasks.

3.4 Introducing missing data

Longitudinal biomedical data almost certainly have missing data. Different missing data mechanisms were introduced in the responses of the complete data mentioned in the first two sections, following the definitions according to Rubin [1976]:

Missing Completely At Random (MCAR):

In this scenario, the assumption is that missing responses are independent of the observed variables. This case does not introduce bias on estimated values. Keeping this in mind, the complete responses Y_{ij} were replaced with missing values with probability p_M, yielding a response with missing values Y_{ij}^M, given by:

$$\begin{aligned} &U_{ij} \sim \mathscr{U}(0,1) \\ &Y_{ij}^M = \begin{cases} \text{NA} & \text{if } U_{ij} \leq p_M \\ Y_{ij} & \text{if } U_{ij} > p_M \end{cases} \end{aligned} \tag{17}$$

Missing At Random (MAR):

In this case, missing responses are only dependent on observed variables. Trying to mantain comparability between different mechanisms, the following criteria was implemented: in order to have a missing response with probability p_M, but dependent of the observed variables in the design matrix $\boldsymbol{X}$ (binding by row all the individual matrices $\boldsymbol{X}_i$), Y_{ij} is replaced by a missing value with probability $p(X_{ij1}, \cdots, X_{ijp})$, where X_{ijk} is the covariable number k, for subject number i,

at measurement occassion j. $p(X_{ij1},\cdots,X_{ijp})$ is attained using weighted sums and logistic functions:

$$\begin{aligned}
&\mu_k = E[X_{ijk}], \quad \sigma_k = \sigma[X_{ijk}] \\
&C = \frac{\log\left(\frac{p_M}{1-p_M}\right)}{\sum_{k=1}^{p}\frac{\mu_k}{\sigma_k}} \\
&K(X_{ij1},\cdots,X_{ijp}) = C\cdot\sum_{k=1}^{p}\frac{X_{ijk}}{\sigma_k} \\
&p(X_{ij1},\cdots,X_{ijp}) = \frac{e^{K(X_{ij1},\cdots,X_{ijp})}}{1+e^{K(X_{ij1},\cdots,X_{ijp})}}
\end{aligned} \tag{18}$$

We note that $E[p(X_{ij1},\cdots,X_{ijp})] = p_M$. Also, the function $K(X_{ij1},\cdots,X_{ijp})$ prioritizes variables with small standard deviations in order to reduce the variability of $p(X_{ij1},\cdots,X_{ijp})$. Thus, Y_{ij}^M is attained replacing p_M by $p(X_{ij1},\cdots,X_{ijp})$ in (17).

Not Missing At Random (NMAR):

In this case, missing responses are dependent on the values of the missing responses. Similar to the MAR proposal, a term was included corresponding to values in Y, with mean μ_Y and standard deviation σ_Y. Therefore, the probability of changing Y_{ij} to a missing value is $p(X_{ij1},\cdots,X_{ijp},Y_{ij})$, where the definitions are similar to those given in (18):

$$\begin{aligned}
&\widetilde{C} = \frac{\log\left(\frac{p_M}{1-p_M}\right)}{\frac{\mu_Y}{\sigma_Y}+\sum_{k=1}^{p}\frac{\mu_k}{\sigma_k}} \\
&\widetilde{K}(X_{ij1},\cdots,X_{ijp},Y_{ij}) = \widetilde{C}\cdot\left(\frac{Y_{ij}}{\sigma_Y}+\sum_{k=1}^{p}\frac{X_{ijk}}{\sigma_k}\right) \\
&p(X_{ij1},\cdots,X_{ijp},Y_{ij}) = \frac{e^{\widetilde{K}(X_{ij1},\cdots,X_{ijp},Y_{ij})}}{1+e^{\widetilde{K}(X_{ij1},\cdots,X_{ijp},Y_{ij})}}
\end{aligned} \tag{19}$$

The response with missing values Y_{ij}^M is obtained replacing p_M by $p(X_{ij1},\cdots,X_{ijp},Y_{ij})$ in (17).

Dealing with missing data

Most procedures dealing with missing data focus on the missing data pattern, which can be thought of as an indicator variable M_{ij} of value 1 if Y_{ij} is observed and 0 otherwise. Assuming structures of zeros and ones in this vector and associated probabilities can improve the estimation of the population parameters.

In some cases, the patterns can be considered monotone (in which if an observation is missing for a certain individual, all subsequent observations of the same individual are missing) or non-monotone (where individuals can have missing observations but observed responses in a posterior timepoint). The former case only admits dropouts from the study and this structure allows the use of specific tools

used for the analysis of the missing data patterns, whereas non-monotone missing data patterns are much more challenging.

Also, most approaches to missing data assume a factorization of the joint likelihood of vectors $\boldsymbol{Y}$ and $\boldsymbol{M}$. Depending on the assumptions that can be made, the missing data pattern $\boldsymbol{M}$ does not influence the estimation of population. This condition is called ignorability and is valid for MCAR and MAR mechanisms, even if the factorizations of the likelihood are different.

However, under NMAR mechanisms, these assumptions cannot be made, the missing data pattern is non-ignorable and the joint likelihood can be factorized according to three main approaches: selection models, pattern-mixture models and shared parameter models (specific for mixed models). Details can be found in Molenberghs et al. [2014].

All these approaches require many assumptions that can be scientifically pertinent, but cannot be verified. Furthermore, since many factors can explain missing data, the assumptions are very specific to the application and the corresponding study. Moreover, in observational studies, many issues are not under the control of researchers and thus, a small number of assumptions can be made. This lack of assumptions often leads researchers to exclude the missing observations and estimate the population parameters dismissing the missing data pattern, at the risk of possible bias. In this work we focus on this case given the impossibility to apply these procedures in general databases without prior knowledge.

3.5 Evaluation of the algorithm

For any given data, the main difficulty in evaluating the performance of the algorithm is the lack of a classification variable indicating which of the response trajectories or measurement occasions are outliers. Therefore, it is of paramount importance to have a reference database in which the detected trajectories are considered as reliable or "true" detections, comparing it to the test database. The simulated databases detailed in Section 3, in which abnormalities are introduced artificially, provide a location and reference for the real abnormalities and the detections can be compared to these values.

Using the simulated databases described in Sections 3.1 and 3.2 and the missing data mechanisms described in Section 3.4, Table 2 shows the comparisons that were considered.

Also, a trajectory or measurement occasion detected as abnormal is considered a "positive detection". Thus, considering $R+$ as a positive detection in the reference database, and $T+$ as a positive detection in the test database, a comparision between the same task in both databases, yields the following contingency

Comparison	Reference	Test
Simulated Database	$\boldsymbol{D}^P$ and $\boldsymbol{D}^R$	SDB
Simulated + Missing Data	$\boldsymbol{D}^P$ and $\boldsymbol{D}^R$	SDB w/MD

Table 2: Comparisons used to evaluate the algorithm. The following abbreviations were used in this table: SDB corresponds a simulated database using $\widehat{\boldsymbol{Y}}_i$ (described in equation 16) as a response variable. Also, MD denotes that missing data was introduced in the response variable. $\boldsymbol{D}^P$ and $\boldsymbol{D}^R$ are defined in Section 3.2.

table:

	$R+$	$R-$
$T+$	TP	FP
$T-$	FN	TN

(20)

where $R-$ and $T-$ are negative detections in the reference and test database, respectively. Furthermore, the values TP represents the number of true positives, and similarly FP, FN, TN are the number of false positives, false negatives and true negatives, respectively. Again, the concept of "true" and "false" are those given by the reference database.

Assesing if the detections in the test databases are "true" or "false" requires tracking the corresponding information of $\boldsymbol{D}^P$ and $\boldsymbol{D}^R$ defined in Section 3.2. For the residual detection task, the true detections are attained identifying the non-zero values of $\boldsymbol{D}^P$, and retrieving the corresponding ID's and timepoints in the simulated database, and comparing them to the detections provided by the algorithm. In a similar way, the non-zero rows of $\boldsymbol{D}^R$ correspond to the ID's with true random effect detections (where the column number identifies the random effect) and can be compared to the detections attained in the output of Algorithm 2.

An important remark is that whenever missing data are introduced, there can be a missing value in the test database in the same location as an object detected by the reference database. Therefore, if the algorithm does not detect an abnormal response because it is missing, it is not counted as a "false negative", since the algorithm is not capable of detecting such event. The same criteria is applied to the missing data that were originally negative and missing in the test database, they are excluded as "true negatives".

These contingency tables yield some quantitative measures that can be used to evaluate the performance of the algorithm:

- **Sensitivity:** $\frac{TP}{TP+FN}$
- **Specificity:** $\frac{TN}{FP+TN}$
- **Positive Predictive Value:** $\frac{TP}{TP+FP}$
- **Negative Predictive Value:** $\frac{TN}{TN+FN}$

For simplicity, from now on the Positive and Negative Predictive Value will be referred as PPV and NPV respectively.

The number of positives in both scenarios is usually low compared to the number of negatives, because the algorithm detects as positive an abnormal behavior. This yields a higher value of TN than all other elements of the contingency table (all other cases are detected as positive in one of both scenarios). Thus, the Specificity and NPV are usually close to 1. On the other hand, sensitivity and PPV can result in a large range of values, since a slight reduction in the numerator can have a considerable impact in the quotient value.

3.6 Results for simulated databases

In the following section the evaluation of the algorithm in each setting and database is presented in several tables and graphs. Keeping in mind the importance of positive detections, the focus is set in the sensitivity and positive predictive value.

Furthermore, if the area of application requires low false negatives the sensitivity may have a higher priority. For example, in medical applications, the cost of missing a positive detection may result in serious consequences, and the health professional may prefer a larger number of detections and rely on their experience to discard false positives.

However, if the PPV is low, the number of false positives is big compared to the detections, and manually discarding false positives can be overwhelming. Therefore, a desirable method would require a balance between values of Sensitivity and PPV.

Another reason why specificity analysis is omitted is that similar results are obtained with all methods and experiments, usually with very high values. Thus, the graphs are not very informative.

Also, the thresholding proposed in Zewotir and Galpin [2007] does not detect any extreme residual in our simulation and therefore, does not figure in the corresponding section.

In all the following sections $M = 100$ repetitions for each parameter combination of the experiment were conducted, in order to have a mean value and standard error to display in the different charts.

Several graphics were used to analyse the results of the detection algorithms in the simulated databases described in Section 3. Since no noticeable difference is observed in performance regarding sample size, in this section, the number of subjects is fixed at $N = 100$. The sample size analyisis is focused on complexity and is developed in Section 3.8.

3.6.1 Extreme Residuals

Figure 3 shows the evolution of the mean sensitivity of the ordinary residuals $\boldsymbol{r}$ based on increasing values of p_P with fixed values of u_P, u_{RI} and u_{RS} and varying p_{RI} and p_{RS}.

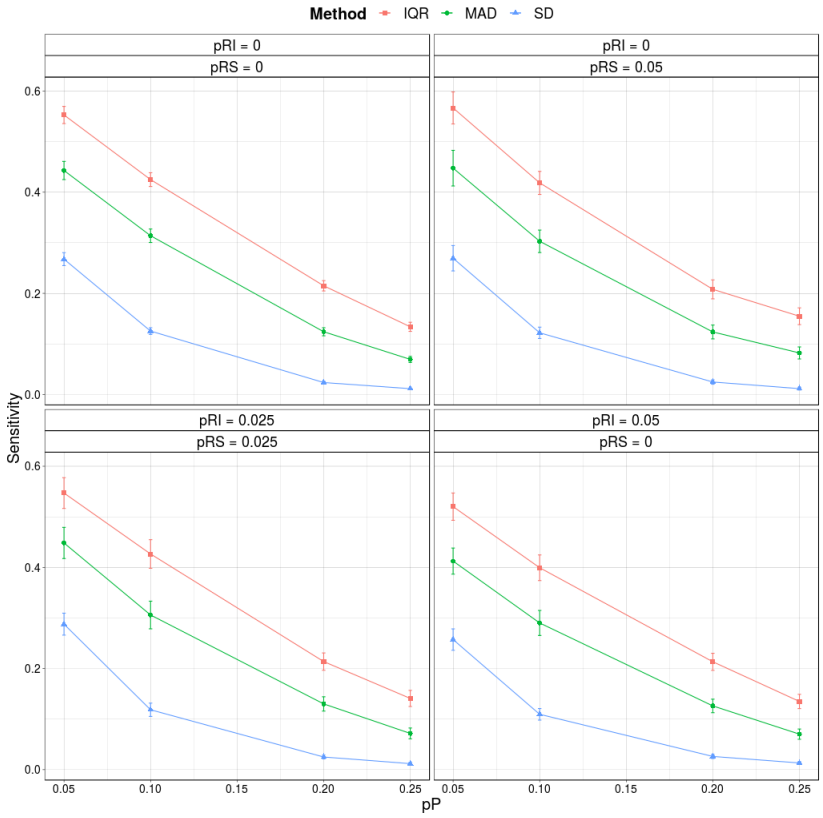


Figure 3: Mean sensitivity and standard error analysis for the residual detection task, with $u_P = 4$ and $u_{RI} = u_{RS} = 1$ as fixed values.

In all panels, the IQR method gives a higher sensitivity. The prevalence of this method in sensitivity analyisis is almost constant throughout the simulations of residual detections because the IQR method is slightly less restrictive than the other methods, yielding a higher rate of positive detection.

It is worth noting that the sensitivity of the SD method decreases faster than the remaining methods, given that the calculation of the standard deviation is more susceptible to noise.

An unexpected result is that when only random slopes are introduced, the sensitivity in the IQR method slightly improves, whereas when only random intercepts are included, the sensitivity decreases its value. This can be explained as

follows: an increase in variability of the random slopes affects all timepoints of the response trajectories, yielding variability in the estimated slopes and therefore, predictions that are more adaptable to the diversity in the responses and therefore, providing in some cases better individual estimations for all timepoints. On the other hand, adding extreme random intercepts to the data only provides estimations that adapt to the variability in the baseline response.

Figure 4 represents the sensitivity analysis for fixed values of p_{RI} p_{RS}, allowing to visualize the difference for different values of u_P, u_{RI} and u_{RS}.

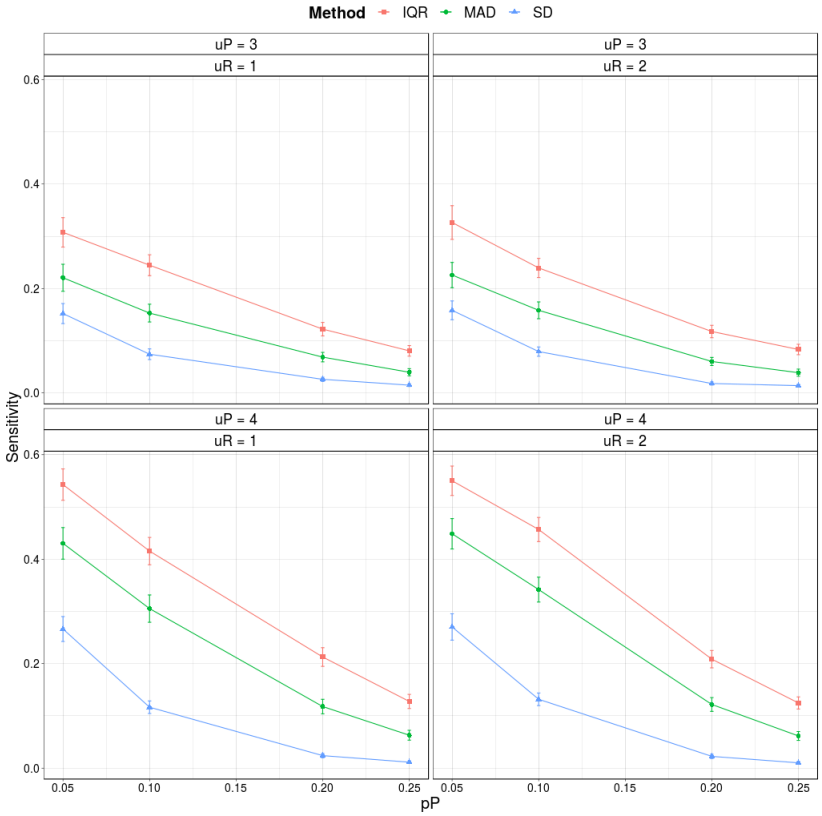


Figure 4: Mean sensitivity and standard error, with $p_{RI} = p_{RS} = 0.025$ as fixed values, for the residual detection task.

As expected, higher values of u_P give rise to bigger peaks, and thus, result in higher sensitivity. It is worth noticing that again the IQR method corresponds to higher sensitivity values. However, there is an unexpected result: keeping u_P constant, an increase in u_R should add noise to the model. However, the IQR method yields a slightly higher sensitivity given this increase. A similar feature is observed for the MAD method with $u_P = 4$.

The PPV analysis for residual detection is shown in Figure 5, with fixed values of p_{RI} and p_{RS} and u_P.

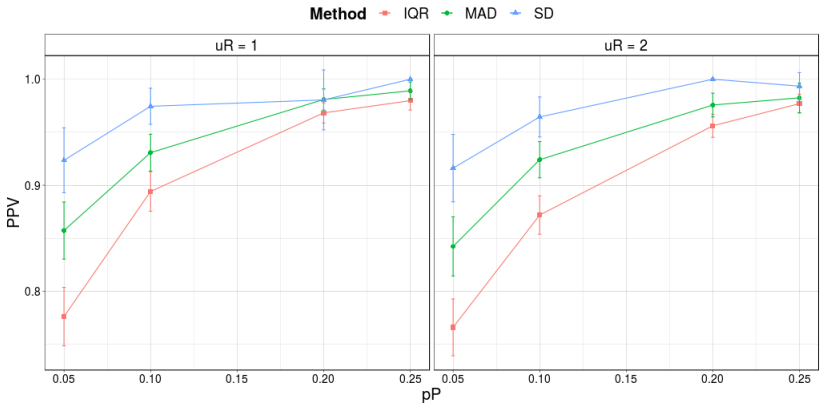


Figure 5: PPV analysis for the residual detection task, keeping as fixed values $p_{RI} = p_{RS} = 0.025$ and $u_P = 4$.

It can be seen in this figure that the SD method yields the highest PPV. Since the SD method is a little more restrictive, there is a lower tolerance to postive detection. Therefore, the positive detections are frequently true positives, yielding a higher PPV in most settings. However, all methods have reasonable PPV values and the SD method usually has a higher false negative rate.

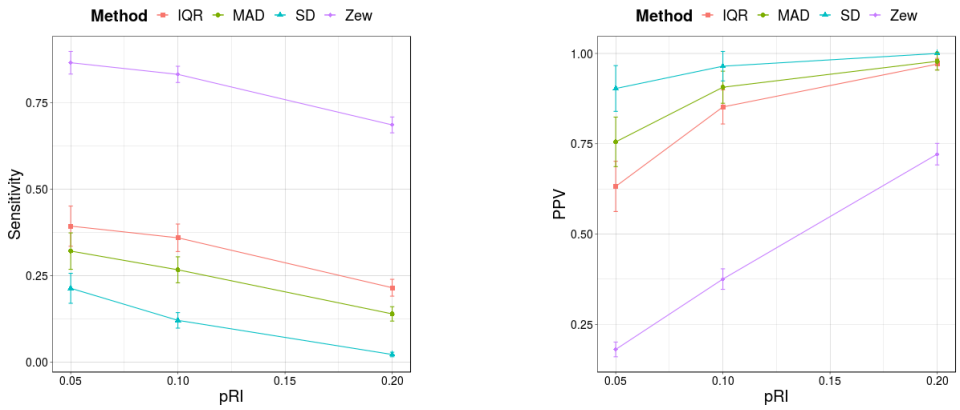
Also, the values of sensitivity using the ordinary, standarized and predicted residuals ($\boldsymbol{r}$, $\boldsymbol{t}$ and $\boldsymbol{t}^*$, respectively) do not show a noticeable difference.

An observed feature in these figures, is that adding noise to the model has no significant effect on the sensitivity. For example, in residual detection, adding extreme random effects ($p_{RI}, p_{RS} > 0$) and with higher magnitude (greater values $u_R = 1$ vs. $u_R = 2$) the only difference in changing these parameters is a slightly higher standard error.

Another noticeable characteristic is that the mean value shows similar morphology and respect a certain order throughout different methods. Keeping this in mind, the following figures will only show some fixed values for the parameters, allowing to visualize the main characteristics.

3.6.2 Extreme Random Intercept

Figure 6 (a) and (b) show the sensitivity and PPV analysis, respectively, for random intercept detection with fixed values of p_P, p_{RS}, u_P and u_R.



(*a*) (*b*)

Figure 6: (*a*) Sensitivity and (*b*) PPV analysis for the random intercept detection task, keeping fixed $p_P = p_{RS} = 0$, $u_P = u_{RS} = 1$ and $u_{RI} = 4$.

Clearly, the thresholding method given by Zewotir and Galpin [2007] yields the highest sensitivity, but at a cost of a very small PPV (only 20% of the detections are true positives when $p_{RI} = 0.05$). The remaining methods have a similar behaviour to the results from the residual detection task, with the IQR yielding a higher sensitivity, compared to the MAD and SD methods.

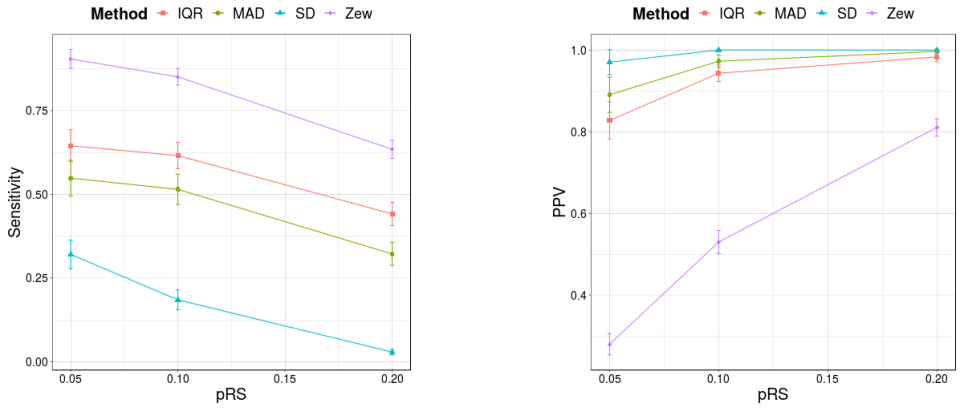
Also, no major differences were observed using the estimated random intercepts ($\widehat{\boldsymbol{b}}_{i1}$) and corresponding the standarization ($\widehat{\boldsymbol{v}}_{i1}$).

3.6.3 Extreme Random Slope

The results for sensitivity in the random slope detection are shown in Figure 7. The Zewotir and Galpin [2007] thresholding uses the standarized random slopes ($\widehat{\boldsymbol{v}}_{i2}$), whereas all other methods use the ordinary estimated random slopes ($\widehat{\boldsymbol{b}}_{i2}$).

In comparison to Figure 6 (*a*), all methods improve their sensitivity and PPV in Figure 7 (*a*) except for the thresholding proposed in Zewotir and Galpin [2007]. This feature is explained by the constant thresholding and the proposed standarization of the estimated random effects, which normalize all random effects to a single distribution. This improvement compared to the random intercept detection can be explained as follows: an extreme random slope has an impact in all timepoints of the response trajectory, whereas a random intercept only affects the baseline, making it harder to detect the impact of the random effect.

Furthermore, the thresholding proposed by Zewotir and Galpin [2007] has a small PPV value. Another noticeable feature is the fast decay of the sensitivity for

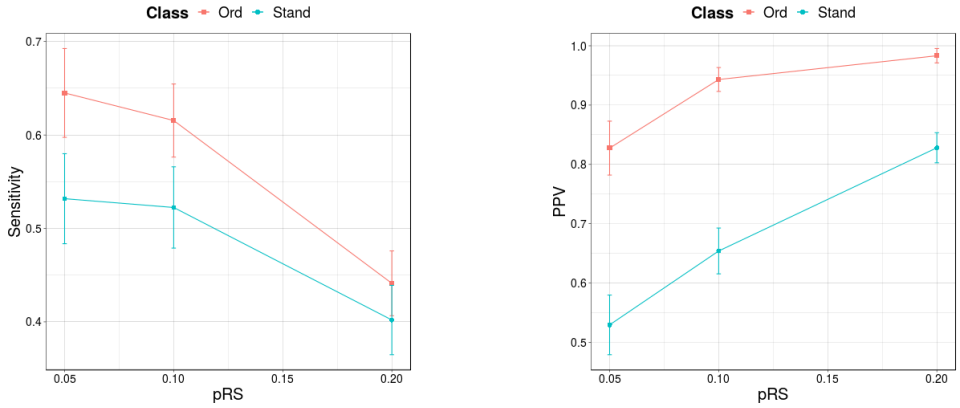


(*a*) (*b*)

Figure 7: (*a*) Sensitivity and (*b*) PPV analysis, keeping fixed $p_P = p_{RI} = 0$, $u_P = u_{RI} = 1$ and $u_{RS} = 4$.

the SD method in comparison to the other methods.

Also, Figure 8 shows how the IQR method performs better using the ordinary random slopes ($\widehat{\boldsymbol{b}}_{i2}$), compared to the standarization ($\widehat{\boldsymbol{v}}_{i2}$).



(*a*) (*b*)

Figure 8: (*a*) Sensitivity analysis, keeping fixed $p_{RI} = p_{RS} = 0$, for the random slope detection task. (*b*) PPV analysis for the random slope detection task, with $p_P = p_{RS} = 0$.

3.7 Results for simulated databases with Missing Data

The Missing Data Mechanisms described in Section 3.4 were applied to the Simulated Databases from the previous section. The analysis was conducted fixing some parameter values, varying the Missing Data Mechanisms (MCAR, MAR and NMAR) and p_M taking the following values: 0.05, 0.1 and 0.2. The results are shown in Figures 9 and 10, yielding the sensitivity values corresponding to the residual detection task, keeping $N = 100, u_P = 4, u_R = 1$ and $p_{RI} = p_{RS} = 0$.

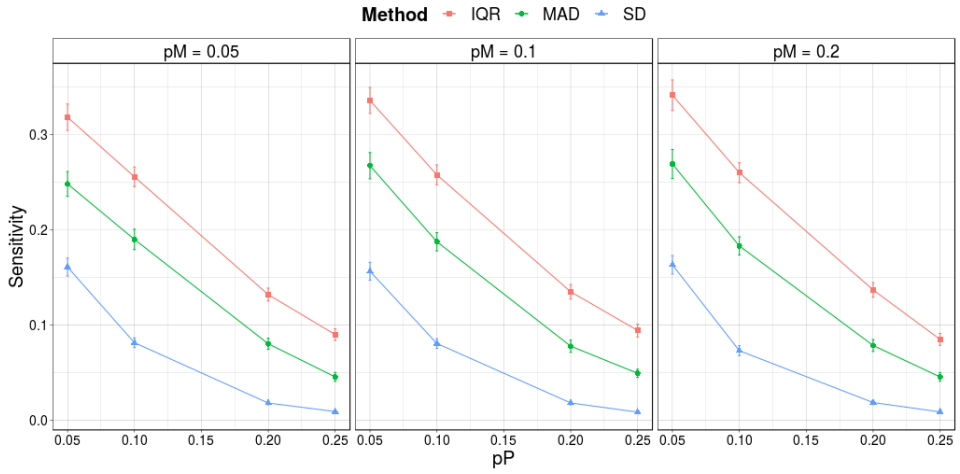


Figure 9: Sensitivity analysis for the residual detection task, under the NMAR missing mechanism.

Excluding data has an effect on all evaluation measures given in Section 3.5, due to changes in the numerator and denominator. However, the morphological tendency of the different evaluation measures are similar to those in Section 3.6 with the same parameter values and detection tasks.

An unexpected result that can be seen in Figure 9 is that a higher proportion of missing values slightly increases the sensitivity.

Also, Figure 10 shows another peculiarity. The sensitivity for the IQR and MAD method is lower in the MCAR mechanism, which is the most desirable setting, given that it does not introduce bias. It seems that the bias of the MAR and NMAR is favoring the detection of extreme residuals, even if the peaks are positive and negative in equal proportions.

Furthermore, for individual random effects detection, an individual is considered missing if the entire response trajectory is missing. Since this event is less frequent, there is no discernable difference in these values after introducing missing data.

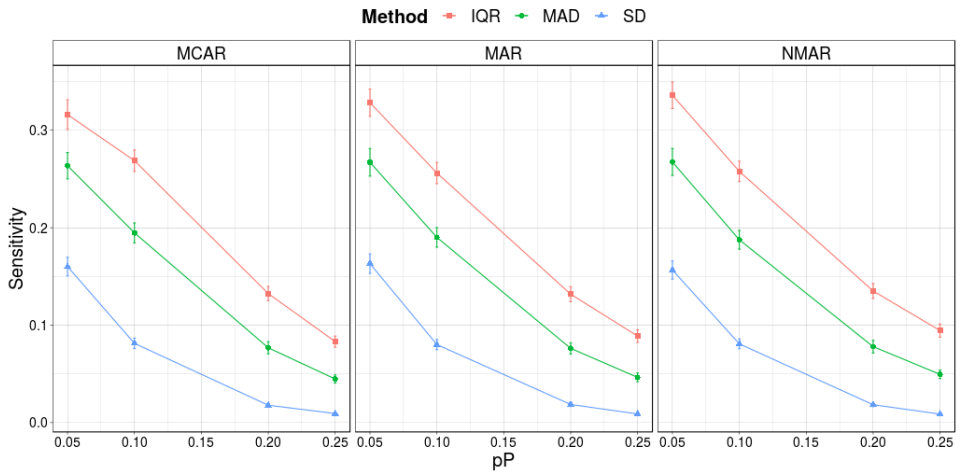


Figure 10: Sensitivity analysis for the residual detection task, keeping p_M =0.1.

3.8 Sample size

Table 3 shows the results for computing time in microseconds of each thresholding option. There is a clear computational difference between using ordinary residuals and random effects, against the corresponding standarizations. This difference can be explained by the different design matrices and the respective inversions involved in the standarizations. Given the block structure of design matrices in longitudinal data, the calculations can be separated per subject. However, the corresponding dimensionality reduction still requires considerable computing time.

All algorithms seem to evolve proportionally to the sample size, except for the Zewotir and Galpin [2007] thresholding, which increases faster than the other methods.

Another issue worth mentioning is that no noticeable differences regarding sample size are observed in the performance measures detailed in 3.5. Therefore, the analyisis is omitted from this work to avoid the inclusion of uninformative figures and tables.

4 Benchmark databases

The proposed algorithms were also tested on several benchmark longitudinal databases. Some are benchmark databases[1] described in Fitzmaurice et al. [2012].

[1] available in https://content.sph.harvard.edu/fitzmaur/ala2e/

Method	Detection	Class	N=100	N=200	N=300
IQR	Residuals	Ordinary ($\boldsymbol{r}$)	123.68(6.652)	201.999(6.172)	304.836(5.004)
		Standarized ($\boldsymbol{t}$)	800.39(6.83)	1540.423(6.001)	2298.749(5.982)
		Predicted ($\boldsymbol{t}^*$)	798.981(6.914)	1542.468(5.414)	2303.471(5.445)
	Random Intercept	Ordinary ($\boldsymbol{B}$)	117.96(4.79)	198.392(2.111)	290.204(4.273)
		Standarized ($\boldsymbol{V}$)	793.246(5.144)	1541.703(3.665)	2285.728(5.288)
	Random Slope	Ordinary ($\boldsymbol{B}$)	130.528(7.727)	200.202(6.058)	286.727(4.165)
		Standarized ($\boldsymbol{V}$)	803.665(7.971)	1524.815(6.079)	2284.487(4.943)
MAD	Residuals	Ordinary ($\boldsymbol{r}$)	123.32(6.688)	196.087(5.055)	289.112(4.381)
		Standarized ($\boldsymbol{t}$)	801.081(7.072)	1545.474(6.253)	2313.759(5.734)
		Predicted ($\boldsymbol{t}^*$)	800.261(7.22)	1544.727(5.825)	2316.206(5.727)
	Random Intercept	Ordinary ($\boldsymbol{B}$)	117.17(4.747)	199.374(2.385)	287.683(3.9)
		Standarized ($\boldsymbol{V}$)	794.22(5.773)	1548.236(3.852)	2310.719(5.12)
	Random Slope	Ordinary ($\boldsymbol{B}$)	129.908(7.717)	198.605(5.902)	284.058(3.686)
		Standarized ($\boldsymbol{V}$)	805.085(8.209)	1543.607(6.577)	2307.901(5.173)
SD	Residuals	Ordinary ($\boldsymbol{r}$)	123.324(6.644)	195.668(5.127)	288.259(4.412)
		Standarized ($\boldsymbol{t}$)	798.961(7.501)	1547.536(5.855)	2326.721(5.995)
		Predicted ($\boldsymbol{t}^*$)	800.217(6.836)	1549.968(5.782)	2319.264(6.356)
	Random Intercept	Ordinary ($\boldsymbol{B}$)	116.627(4.716)	197.859(2.079)	286.553(3.901)
		Standarized ($\boldsymbol{V}$)	795.396(5.858)	1542.785(3.578)	2308.299(5.068)
	Random Slope	Ordinary ($\boldsymbol{B}$)	129.851(7.688)	198.693(5.879)	283.707(3.734)
		Standarized ($\boldsymbol{V}$)	807.589(8.283)	1548.965(6.198)	2317.318(5.556)
Zew	Residuals	Standarized ($\boldsymbol{t}$)	824.169(7.106)	1794.68(7.252)	3350.533(10.55)
	Random Intercept	Standarized ($\boldsymbol{V}$)	817.616(5.238)	1786.657(4.915)	3325.185(8.747)
	Random Slope	Standarized ($\boldsymbol{V}$)	829.924(8.372)	1809.246(7.23)	3345.491(8.43)

Table 3: Mean computing time and standard error (in microseconds) for each thresholding of the different detection tasks, for different sample sizes

4.1 FEV_1 Data

The FEV_1 database corresponds to a large longitudinal study (Dockery et al. [1983]) designed to analize evolution of respiratory attributes of children from 6 cities of the US. The original study had 13.379 enrollments. However, the available data consists of N = 300 female children randomly selected from the study subjects coming from Topeka (Kansas). The collected attributes are the age of the participant, the height and the Forced Expiratory Volume during 1 second (noted FEV_1, a measure of pulmonary function). The measurements were repeated annually. However, due to attrition or mistimed measurements, the data is unbalanced over time.

The response trajectories are shown in Figure 11. Fitzmaurice et al. [2012] propose the following mixed effects model to represent the responses (FEV_1) of subject number *i* over time:

$$\begin{aligned} \log(\mathbf{FEV_{1}}_{,i}) \;=\; & \beta_0 \cdot \mathbf{1} + \beta_1 \cdot \mathbf{age}_i + \beta_2 \cdot \log(\mathbf{Ht}_i) + \beta_3 \cdot \mathbf{age}_i^0 \cdot \mathbf{1} + \\ & \beta_4 \cdot \log(\mathbf{Ht}_i^0) \cdot \mathbf{1} + b_{i0} \cdot \mathbf{1} + b_{i1} \cdot \mathbf{age}_i + \boldsymbol{\varepsilon}_i \end{aligned} \tag{21}$$

where $\mathbf{age}_i^0$ and $\mathbf{Ht}_i^0$ stand for the baseline age and height for subject *i*, respectively,

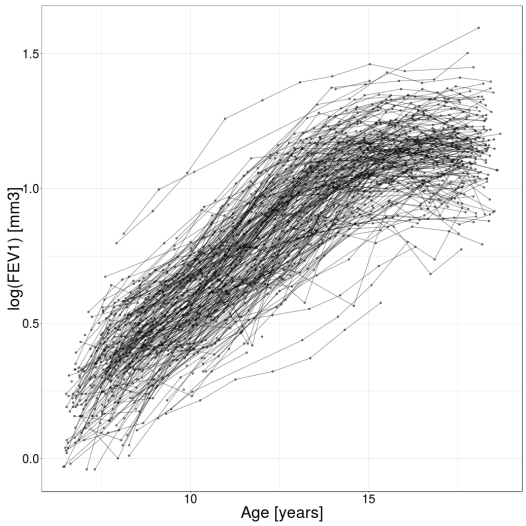


Figure 11: Response trajectories for N=300 randomly selected children from the Six Cities Study, from Topeka (Kansas)

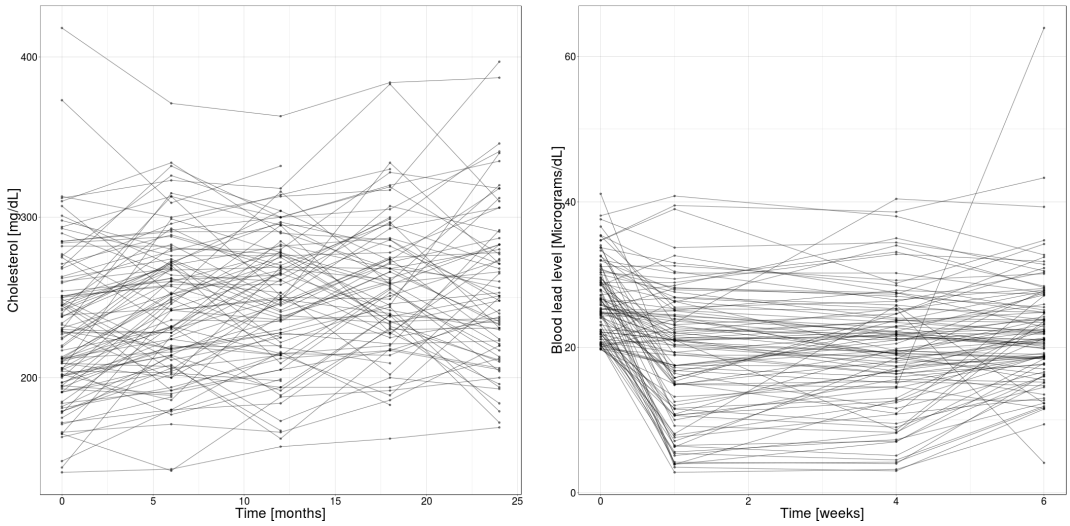
and $\mathbf{1}$ represents a $n_i \times 1$ vector of ones used to account for the same intercept through repeated measurements. Moreover, $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$ represent the fixed effects coefficients,whereas $b_{0,i}$ and $b_{1,i}$ represent the individual random intercept and slope respectively for subject i.

4.2 Cholesterol Data

The Cholesterol Data comes from a study (Wei and Lachin [1984]) conducted to investigate the effects and safety of a drug named chenodiol to treat cholesterol gallstones applied to 103 patients. Two groups were randomly assigned to treatment or placebo. The response variable (Serum cholesterol, measured in mg/dL) was measured at baseline and at 6, 12, 20, and 24 months of follow-up. 68 measurements of the study are missing due to various reasons, yielding a total of N=447 responses.

The response trajectories can be seen in Figure 12 and the corresponding mixed effects model is

$$\mathbf{Chol}_i = \beta_0 \cdot \mathbf{1} + \beta_1 \cdot \mathbf{Time}_i + b_{i0} \cdot \mathbf{1} + b_{i1} \cdot \mathbf{Time}_i + \boldsymbol{\varepsilon}_i \tag{22}$$



(*a*) (*b*)

Figure 12: Response trajectories for subjects of (*a*) the National Cooperative Gallstone Study and (*b*) the Treatment of Lead-Exposed Children Trial.

4.3 TLC Data

The Treatment of Lead-Exposed Children (TLC) trial (Rogan et al. [2000]) was a randomized study to analyse the effects of a drug named succimer in children with similar blood lead levels. These data consist of four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4, and week 6 on 100 children, randomly assigned to treatment with succimer or placebo, yielding a total of N=400 responses.

The response trajectories are shown in Figure 12 and the corresponding mixed effects model is

$$\begin{aligned} \mathbf{Lead}_i &= \beta_0 \cdot \mathbf{1} + \beta_1 \cdot \mathbf{Time}_i + \beta_3 \cdot \mathbf{Trt}_i \cdot \mathbf{Time}_i + \beta_4 \cdot \mathbf{Trt}_i \cdot \mathbf{Time*}_i + \\ & \quad b_{i1} \cdot \mathbf{Time}_i + b_{i2} \cdot \mathbf{Time*}_i + \boldsymbol{\varepsilon}_i \end{aligned} \tag{23}$$

where variable $\mathbf{Trt}_i$ counts as a vector of ones if subject i corresponds to the Succimer group and 0 otherwise. Also, the variable named Time* is defined as:

$$\text{Time*} = \max\{0; \text{Time} - 1\}, \tag{24}$$

simbolyzing time elapsed since the first week of the study.

4.4 Performance in Benchmark databases

Table 4 details the number of outliers found by each thresholding method for the databases described in Section 4. For the thresholding proposed in Zewotir and Galpin [2007], some values are blank since the method uses only one of the options for both residual and random effect detections.

Data	Detection	Class	IQR	MAD	SD	Zew
$\mathbf{FEV_1}$	**Residuals**	**Ordinary** ($\boldsymbol{r}$)	39	30	25	–
		Standarized ($\boldsymbol{t}$)	41	28	25	0
		Predicted ($\boldsymbol{t}^*$)	41	28	25	–
	Random Effects	**Ordinary** ($\boldsymbol{B}$)	22	10	4	–
		Standarized ($\boldsymbol{V}$)	6	2	2	522
Cholesterol	**Residuals**	**Ordinary** ($\boldsymbol{r}$)	4	0	0	–
		Standarized ($\boldsymbol{t}$)	4	1	1	417
		Predicted ($\boldsymbol{t}^*$)	14	13	4	–
	Random Effects	**Ordinary** ($\boldsymbol{B}$)	4	3	2	–
		Standarized ($\boldsymbol{V}$)	4	3	2	0
TLC	**Residuals**	**Ordinary** ($\boldsymbol{r}$)	19	19	5	–
		Standarized ($\boldsymbol{t}$)	21	21	6	208
		Predicted ($\boldsymbol{t}^*$)	23	22	6	–
	Random Effects	**Ordinary** ($\boldsymbol{B}$)	4	6	0	–
		Standarized ($\boldsymbol{V}$)	4	6	0	0

Table 4: Number of outliers detected in the Benchmark databases

For the FEV_1 data, given each method, there is no difference regarding the use of ordinary, standarized or predicted residuals. However, the difference is in the random effect detection, where the standarized random effects yields less detections than the ordinary random effects. Also, the Zewotir and Galpin [2007] thresholding does not detect any extreme residuals, yet classifies a great number of random effects as extreme, given the total number ($N \times q = 600$). Both results are consistent with those obtained in the simulations, i.e., low PPV in the random effect detection task.

The opposite situation is observed in both Cholesterol and TLC data: the standarization of the random effects does not affect the number of detections, whereas there is a difference regarding ordinary, standarized and predicted residuals. Also, the Zewotir and Galpin [2007] thresholding yields a great number of residual detections, compared to the number of total observations.

The erratic results for the Zewotir and Galpin [2007] thresholding can be explained by basing great part of the constant boundaries on the sample size, whereas the other methods rely on the dispersion of the corresponding vector. The constant

thresholding can be extremely restrictive or tolerant, and does not adapt well to different types of data and thus, the number of detections do not seem reasonable in these databases.

Furthermore, the differences between FEV_1 data and the rest can be explained by the sample size: FEV_1 data has $N_T = 1993$ observations, whereas Cholesterol and TLC data have $N_T = 447$ and $N_T = 400$ observations, respectively. Regarding the difference in the random effect detection in FEV_1 data, the same standarization is applied to all random effects without considering possible differences between them. When the number of estimated random effects is large, considering the difference between each column of matrix $\boldsymbol{B}$ becomes more noticeable. Also, the smaller sample size in Cholesterol and TLC data allows each observation to have a greater influence on the estimations. Therefore, removing each observation for the predicted residual calculation can have a greater impact on the vector.

5 Discussion

A new approach for the simultaneous detection of contextual and collective anomalies was presented. The corresponding approach is based on linear mixed effects models, and resorts to different dispersion measures described previously in the literature in other contexts.

The algorithm performs quickly and with a good balance between Sensitivity and PPV applying the IQR method to the ordinary residuals and estimated random effects. Although the Zewotir and Galpin [2007] thresholding (included for the sake of comparison) yields a higher sensitivity than the IQR method for the random effect detection task, the number of false positives is too high for databases with relatively large sample sizes.

Furthermore, whenever the boundaries are based on a dispersion measure, such as in the IQR method, the results adapt to the variability of the data at hand. On the other hand, thresholds mainly established by the sample size can yield too restrictive or too tolerant cutoff points.

Also, for boundaries that rely on specific calculations for a certain statistical model, the performance of the corresponding algorithm decreases if the data does not adjust precisely to the model. Since in real observational data model assumptions ususally are not totally met, the IQR method can provide satisfactory results in these cases.

In addition, different random effects have different influences on the response trajectories. Therefore, considering separate detections for each variable allows to profit from the different interpretations of each random effect, without recurring to a normalization that applies uniformly to all random coefficients.

Regarding alternatives to some features in this work, we can mention the following approaches:

The algorithm does not require the value of $T = 1.5$ for the IQR Method or $T = 3$ for the MAD and SD Methods. These tolerance values can be changed to achieve a better performance of the detection tasks.

For example, the boundaries can be based on the application. The user may prefer to detect an absolute or percentual difference between the empirical values and the fitted values. Furthermore, the observations can be ranked according to their distance to the null value, allowing to manually adjust the percentage of observations with the highest absolute value to the application's need.

Moreover, if the users can use their expertise to deem which of the detections are true or false positives, the algorithm can improve if the false positives share similar characteristics (suggesting a confounding variable) or the established boundary was excesively tolerant.

Different methods can be used for each detection task, depending on the priorities established by the user. Since each detection task has a respective univariate vector with a corresponding interpretation, many of the cluster-based or density-based methods described in Section 1 can be applied to a residual vector or a column of the matrix of estimated random effects.

As it was mentioned in Section 1, the Mahalanobis distance does not apply to vectors of different length, given that there is not a shared covariance matrix and hence, methods based on this distance could not be applied to the response trajectories when there are missing observations. However, by using Equation (5) an individual covariance matrix can be estimated for each response trajectory, and the Mahalanobis distance can be used. Furthermore, every strategy can be improved using robust estimates for the parameters, at the cost of an increased computational complexity.

Modifications in some features in Section 3 can be introduced. For example, to achieve mistimed measurements, the measurement occasions were determined using random exponential variables. However, these timepoints can be attained by using other non-negative random variables such as gamma, weibull or simply non-negative uniform distributions, instead of exponential variables.

The missing data mechanisms described in 3.4 use weighted sums that guarantee equal expected probabilities of obtaining missing values for the proposed MCAR, MAR and NMAR mechanisms. These weights may take other values as long as the expected proportion of missing data agrees with a specified probabilty.

The algorithm may improve by adding another step in the estimation of the individual response. The current algorithm estimates the mixed model parameters by excluding the missing responses. These prior estimations can be used to perform a model-based multiple imputation of the responses and to estimate the model

parameters once more.

For the simulated data with missing responses described in 3.4, the missing data pattern is known. Therefore, the likelihood-based methods discussed in 3.4 can be applied to improve the estimation of the model parameters with verifiable hypotheses. With the aim of applying these extensions to real data, the performance of different thresholding methods can be compared whenever assumptions regarding the pattern of missing values are reasonable.

References

Abraham, B. and G. E. Box (1979): "Bayesian analysis of some outlier problems in time series," *Biometrika*, 66, 229–236.

Abraham, B. and A. Chuang (1989): "Outlier detection and time series modeling," *Technometrics*, 31, 241–248.

Beckman, R. J. and R. D. Cook (1983): "Outlier. s," *Technometrics*, 25, 119–149.

Bellazzi, R., M. Diomidous, I. N. Sarkar, K. Takabayashi, A. Ziegler, and A. T. McCray (2011): "Data analysis and data mining: current issues in biomedical informatics," *Methods of information in medicine*, 50, 536–544.

Bianco, A. M., M. Garcia Ben, E. Martinez, and V. J. Yohai (2001): "Outlier detection in regression models with arima errors using robust estimates," *Journal of Forecasting*, 20, 565–579.

Billor, N., A. S. Hadi, and P. F. Velleman (2000): "Bacon: blocked adaptive computationally efficient outlier nominators," *Computational statistics & data analysis*, 34, 279–298.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, and J. Sander (2000): "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 93–104.

Chandola, V., A. Banerjee, and V. Kumar (2009): "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, 41, 1–58.

Chatterjee, S., A. S. Hadi, et al. (1986): "Influential observations, high leverage points, and outliers in linear regression," *Statistical science*, 1, 379–393.

Chawla, N. V. and D. A. Davis (2013): "Bringing big data to personalized healthcare: a patient-centered framework," *Journal of general internal medicine*, 28, 660–665.

Cowie, M. R., J. I. Blomster, L. H. Curtis, S. Duclaux, I. Ford, F. Fritz, S. Goldman, S. Janmohamed, J. Kreuzer, M. Leenay, et al. (2017): "Electronic health records to facilitate clinical research," *Clinical Research in Cardiology*, 106, 1–9.

Davies, L. and U. Gather (1993): "The identification of multiple outliers," *Journal of the American Statistical Association*, 88, 782–792.

Delannay, N., C. Archambeau, and M. Verleysen (2008): "Improving the robustness to outliers of mixtures of probabilistic pcas," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 527–535.

Dockery, D., C. Berkey, J. Ware, F. Speizer, and B. Ferris Jr (1983): "Distribution of forced vital capacity and forced expiratory volume in one second in children 6 to 11 years of age," *American Review of Respiratory Disease*, 128, 405–412.

Doukas, C., T. Pliakas, and I. Maglogiannis (2010): "Mobile healthcare information management utilizing cloud computing and android os," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, IEEE, 1037–1040.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. (1996): "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, volume 96, volume 96, 226–231.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2012): *Applied longitudinal analysis*, volume 998, John Wiley & Sons.

Fox, A. J. (1972): "Outliers in time series," *Journal of the Royal Statistical Society: Series B (Methodological)*, 34, 350–363.

Freund, R. J., R. C. Littell, and P. C. Spector (1986): *SAS system for linear models.*, Statistical Analysis System Institute, Incorporated.

Hansen, M., T. Miron-Shatz, A. Lau, and C. Paton (2014): "Big data in science and healthcare: a review of recent literature and perspectives," *Yearbook of medical informatics*, 23, 21–26.

Hardin, J. and D. M. Rocke (2004): "Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator," *Computational Statistics & Data Analysis*, 44, 625–638.

Iglewicz, B. (2014): "Boxplot," *Wiley StatsRef: Statistics Reference Online*.

Kannan, K. S., K. Manoj, and S. Arumugam (2015): "Labeling methods for identifying outliers," *International Journal of Statistics and Systems*, 10, 231–238.

Koh, H. C., G. Tan, et al. (2011): "Data mining applications in healthcare," *Journal of healthcare information management*, 19, 65.

Kriegel, H.-P., P. Kröger, E. Schubert, and A. Zimek (2008): "A general framework for increasing the robustness of pca-based correlation clustering algorithms," in *International Conference on Scientific and Statistical Database Management*, Springer, 418–435.

Lau, F., M. Price, J. Boyd, C. Partridge, H. Bell, and R. Raworth (2012): "Impact of electronic medical record on physician practice in office settings: a systematic review," *BMC medical informatics and decision making*, 12, 10.

Leroy, A. M. and P. J. Rousseeuw (1987): "Robust regression and outlier detection," *Wiley Series in Probability and Mathematical Statistics*.

Lin, J., E. Keogh, A. Fu, and H. Van Herle (2005): "Approximations to magic:

Finding unusual medical time series," in *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, IEEE, 329–334.

Liu, B., Y. Xiao, L. Cao, Z. Hao, and F. Deng (2013): "Svdd-based outlier detection on uncertain data," *Knowledge and information systems*, 34, 597–618.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke (2014): *Handbook of missing data methodology*, CRC Press.

Newman, D. A. (2003): "Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques," *Organizational Research Methods*, 6, 328–362.

Peek, N., J. H. Holmes, and J. Sun (2014): "Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics," *Yearbook of medical informatics*, 23, 42–47.

Ramaswamy, S., R. Rastogi, and K. Shim (2000): "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 427–438.

Roberts, S. J. (2000): "Extreme value statistics for novelty detection in biomedical data processing," *IEE Proceedings-Science, Measurement and Technology*, 147, 363–367.

Rogan, W., R. Bornschein, J. Chisolm, A. Damokosh, D. Dockery, M. Fay, R. Jones, G. Rhoads, N. Ragan, M. Salganik, et al. (2000): "Safety and efficacy of succimer in toddlers with blood lead levels of 20-44 μg/dl," *Pediatric Research*, 48, 593–599.

Rousseeuw, P. J. and B. C. Van Zomeren (1990): "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical association*, 85, 633–639.

Rubin, D. B. (1976): "Inference and missing data," *Biometrika*, 63, 581–592.

Schubert, E., A. Zimek, and H.-P. Kriegel (2014): "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data mining and knowledge discovery*, 28, 190–237.

Sim, C. H., F. F. Gan, and T. C. Chang (2005): "Outlier labeling with boxplot procedures," *Journal of the American Statistical Association*, 100, 642–652.

Suling, M. and I. Pigeot (2012): "Signal detection and monitoring based on longitudinal healthcare data," *Pharmaceutics*, 4, 607–640.

Tsay, R. S., D. Pena, and A. E. Pankratz (2000): "Outliers in multivariate time series," *Biometrika*, 87, 789–804.

Verbeke, G. and E. Lesaffre (1997): "The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data," *Computational Statistics & Data Analysis*, 23, 541–556.

Verbeke, G. and G. Molenberghs (2000): "A model for longitudinal data," *Linear mixed models for longitudinal data*, 19–29.

Wei, L. and J. Lachin (1984): "Two-sample asymptotically distribution-free tests for incomplete multivariate observations," *Journal of the American Statistical Association*, 79, 653–661.

Yoo, I., P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua (2012): "Data mining in healthcare and biomedicine: a survey of the literature," *Journal of medical systems*, 36, 2431–2448.

Zewotir, T. and J. S. Galpin (2007): "A unified approach on residuals, leverages and outliers in the linear mixed model," *Test*, 16, 58–75.

Zhang, D. and M. Davidian (2001): "Linear mixed models with flexible distributions of random effects for longitudinal data," *Biometrics*, 57, 795–802.