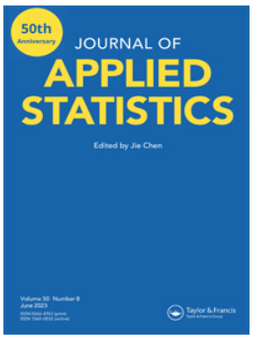

Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Conformal normal curvature and detection of masked observations in multivariate null intercept measurement error models

Reiko Aoki, Juan P. Mamani Bustamante, Cibele M. Russo & Gilberto A. Paula

To cite this article: Reiko Aoki, Juan P. Mamani Bustamante, Cibele M. Russo & Gilberto A. Paula (2023): Conformal normal curvature and detection of masked observations in multivariate null intercept measurement error models, Journal of Applied Statistics, DOI: 10.1080/02664763.2023.2212332

To link to this article: https://doi.org/10.1080/02664763.2023.2212332

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2023.2212332

Taylor & Francis
Taylor & Francis Group

Conformal normal curvature and detection of masked observations in multivariate null intercept measurement error models

Reiko Aoki[a], Juan P. Mamani Bustamante[a], Cibele M. Russo[a] and Gilberto A. Paula[b]

[a]Instituto de Ciências Matemáticas e de Computaç ao, Universidade de São Paulo, São Carlos, Brazil;
[b]Instituto de Matemática e Estatística, Universidade de São Paulo, São Carlos, Brazil

ABSTRACT

Measurement errors occur very commonly in practice. After fitting the model, influence diagnostics is an important step in statistical data analysis. The most frequently used diagnostic method for measurement error models is the local influence. However, this methodology may fail to detect masked influential observations. To overcome this limitation, we propose the use of the conformal normal curvature with the forward search algorithm. The results are presented through easy to interpret plots considering different perturbation schemes. The proposed methodology is illustrated with three real data sets and one simulated data set, two of which have been previously analyzed in the literature. The third data set deals with the stability of the hygroscopic solid dosage in pharmaceutical processes to ensure the maintenance of product safety quality. In this application, the analytical mass balance is subject to measurement errors, which require attention in the modeling process and diagnostic analysis.

ARTICLE HISTORY
Received 13 January 2021
Accepted 4 May 2023

KEYWORDS
Local influence diagnostics; structural relationship; forward search; error in variables model; latent variable; likelihood displacement

1. Introduction

Measurement error models have been applied in many areas of science and received much attention in the past decades (see Refs. [10,14] and references therein). There are plenty of applications where the covariate is not measured precisely such as blood pressure, degree of pest infestation, dental plaque index, intelligence quotient, temperature, mass of a substance and so forth. More recently, Carroll *et al.* [11] studied the prediction problem in non-parametric measurement error models, whereas Zhang *et al.* [39] introduced the linear model selection when the covariates are measured with error. Hu and Wansbeek [21] presented recent studies on measurement error models in the econometric literature. When the covariate is measured with error, it is well known that the naive estimator of the slope parameter obtained by fitting the usual regression model results in a inconsistent estimator (see Ref. [14]).

CONTACT Reiko Aoki reiko@icmc.usp.br

Supplemental data for this article can be accessed here. https://doi.org/10.1080/02664763.2023.2212332

After fitting a model, the influence diagnostics is an important step in statistical data analysis. The usual methodologies to assess the observations that are influential/outliers are the global influence, where a case is removed to measure the effect of the deletion in the parameter estimates, predictions or tests of interest (see, for example, Refs. [7,13,17,38]). In the linear regression, it is well known that the presence of such observation(s) may alter drastically the analyses of a regression model. The usual measurement error model estimates can also behave very poorly in the presence of such observation(s) (see Ref. [14] and references therein). 'The parameter estimates for the measurement error model are even less robust than the least-squares estimates are for the ordinary regression model, and the latter are known to be quite nonrobust ' [14]. Considering the measurement error models and the influence diagnostics, Kelly [23] proposed an influence function for the structural models, Zhong *et al.* [40] dealt with the assessment of local and global influence for linear measurement error models based upon the corrected likelihood of Nakamura [31], and Rasekh and Fieller [36] derived an influence function in functional measurement error models with replicated data. There is an extensive literature with applications of the local influence of Cook [16] in measurement error models. See, for example, Refs. [3,18,25–29,35].

Cook [16] introduced the local influence of minor perturbations in the data set or in the model to identify a group of observations that may exert an undue influence. When the individual cases are deleted (global influence), it may not identify, for example, a group of observations which are jointly influential but not individually influential (see Refs. [9,15]). The local influence is a simple and powerful method, which is based on the normal curvature to study the behavior of the likelihood displacement function. It was suggested to analyze the direction with the largest normal curvature and the relative sizes of its components to identify observations that are subject to maximal sensitivity to the perturbation. However, there was no objective benchmark to judge largeness. To bypass this difficulty, Poon and Poon [33] proposed the use of the conformal normal curvature and an objective benchmark to judge largeness.

These methodologies start by fitting the model to the whole data set including the outliers, which may cause the masking effects, i.e. when an outlier is not detected because of the presence of a cluster of outliers. To overcome this problem, Atkinson and Riani [5] proposed the forward search in regression models that starts fitting the model to very few observations in a robust way using least squares and residuals. The methodology gradually increments the number of observations used in the fit until all the observations are fitted. The key concept of the forward search algorithm is the ordering of the data on the basis of observational residuals to detect multiple masked outliers. In the forward search, the evolution of residuals, parameter estimates and inferences is monitored as the subset size increases [5].

Many articles have been published considering the forward search in different contexts of applications and theories. Mavridis and Moustaki [30] used the forward search algorithm to identify atypical observations in factor analysis models, while Bellini [8] extended the forward search in elliptical copulas and Atkinson *et al.* [6] extended the forward search algorithm to multivariate data. Cerioli *et al.* [12] studied some asymptotic properties of the forward search, Johansen and Nielsen [22] studied the asymptotic properties of the sequence of regression estimators and forward residuals and Grané *et al.* [19]

combined forward search distance-based algorithm with robust clustering to visualize mixed data.

While there are many methodologies to deal with diagnostic analysis in linear regression models, see also Ref. [4], the same is not true with other models. The novelty in this work is the development of a methodology to detect masked influential observations in measurement error models, considering conformal normal curvature and forward search. As previously stated, in measurement error models, it is very important to detect such observations as otherwise the estimates can behave very poorly. Moreover, once the usual local influence analysis is performed, it is easy to perform the proposed methodology with little extra effort and the procedure may reveal influential observation that was masked and not identified during the local influence analyzes.

To show the versatility of the proposed methodology, we apply it to three real data sets and a simulated data set.

The first and second data sets were previously analyzed in the literature and the third data set is from a stability study of a hygroscopic solid dosage.

Aoki *et al.* [1] analyzed a study designed to test the efficacy of two types of toothbrushes in removing dental plaque. In that study, 26 preschoolers were evaluated under these 2 experimental conditions. As null pretest dental plaque indices imply null expected posttest values, the null intercept model was considered. Also, as the dental plaque indices are evaluated imprecisely, the pretest dental plaque indices, as well as the posttest dental plaque indices (after the use of each toothbrush) must account for the measurement errors. Thus the use of the measurement error model was proposed. Furthermore, the proposed model allowed for correlated individual measurements since each preschooler used both of the toothbrushes. We will refer to this data set as toothbrush data.

Considering the toothbrush data just described and a general setting with p treatments, let $\boldsymbol{z}_j$ denote the observed vector for the jth subject, given by

$$\boldsymbol{z}_j = \begin{pmatrix} \boldsymbol{x}_j \\ \boldsymbol{y}_j \end{pmatrix} \quad \text{with } \boldsymbol{x}_j = (x_{1_j}, \ldots, x_{p_j})^T \quad \text{and} \quad \boldsymbol{y}_j = (y_{1_j}, \ldots, y_{p_j})^T,\ j = 1, \ldots, n.$$

Then, Aoki *et al.* [1] extended the classical measurement error model proposing the following model:

$$\begin{aligned} y_{i_j} &= \beta_i \xi_{i_j} + e_{i_j}; \\ x_{i_j} &= \xi_{i_j} + \delta_{i_j}; \\ \xi_{i_j} &= \mu + a_j; \end{aligned} \tag{1}$$

where $a_j \overset{ind.}{\sim} N(0, \sigma_x^2)$, $\delta_{i_j} \overset{ind.}{\sim} N(0, \sigma_\delta^2)$, $e_{i_j} \overset{ind.}{\sim} N(0, \lambda_i \sigma_\delta^2)$, δ_{i_j}, e_{i_j} and a_j independent for $i = 1, \ldots, p$, $j = 1, \ldots, n$. The term a_j allows for a possible within subject correlation structure, leading to the random effect model. Thus, the observed vector for the jth subject $\boldsymbol{z}_j \sim N_{2p}(\boldsymbol{m}, \boldsymbol{V})$, $j = 1, \ldots, n$, where $\boldsymbol{m} = \mu \boldsymbol{b}$, $\boldsymbol{V} = \sigma_\delta^2 \boldsymbol{A} + \sigma_x^2 \boldsymbol{b}\boldsymbol{b}^T$ with $\boldsymbol{b} = (\boldsymbol{1}_p^T, \boldsymbol{\beta}^T)^T$,

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{I}_P & \boldsymbol{0}_P \\ \boldsymbol{0}_P & \boldsymbol{D}(\boldsymbol{\lambda}) \end{pmatrix},$$

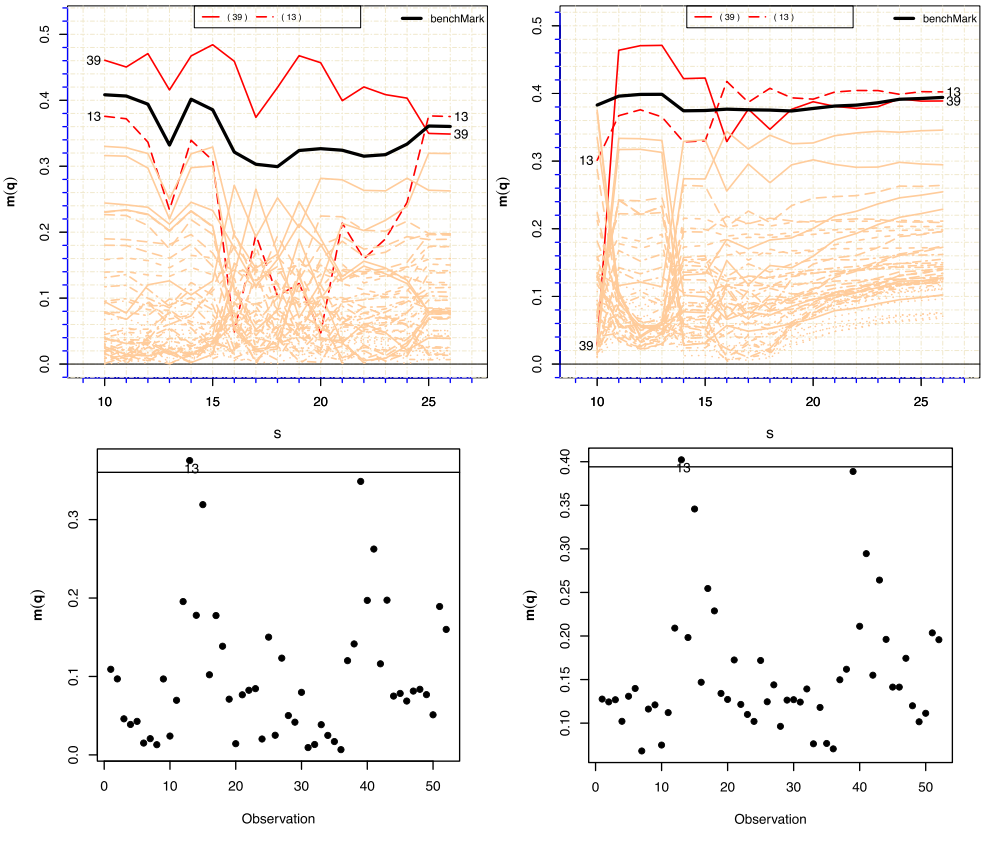

Figure 1. Toothbrush data: explanatory variable perturbation scheme. CNCFS forward plot (top) and index plot of $\boldsymbol{m}(q)$ (bottom) for the contribution of the eigenvector associated with the largest eigenvalue in the left-hand panels and the total contribution ($q = 0$) in the right-hand panels.

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T, \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T$, $\mathbf{1}_p$ denoting a vector composed by p 1's, $\boldsymbol{I}_P$ the identity matrix of order p, $\mathbf{0}_P$ a square matrix of order p composed by 0's and $\boldsymbol{D}(\boldsymbol{\lambda})$ a diagonal matrix with elements of the vector $\boldsymbol{\lambda}$. Furthermore, the log likelihood function is given by

$$\ell(\boldsymbol{\theta}) = -np\log(2\pi) - \frac{n}{2}\log|\boldsymbol{V}| - \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{z}_j - \boldsymbol{m})^T\boldsymbol{V}^{-1}(\boldsymbol{z}_j - \boldsymbol{m}) \tag{2}$$

with $\boldsymbol{\theta}_{(2p+3)\times 1} = (\boldsymbol{\beta}^T, \mu, \sigma_x^2, \sigma_\delta^2, \boldsymbol{\lambda}^T)^T$.

In the case of the toothbrush data, we have $p = 2$ with $i = 1$ representing the experimental toothbrush and $i = 2$ representing the conventional toothbrush. The number of preschoolers in that study was $n = 26$. The Anderson–Darling (AD) and Cramér–von Mises (CVM) tests were considered to test for normality (see Refs. [24,32]). The p-value for the AD test was given by 0.837 and for the CVM test was given by 0.937. Consequently, considering the significance level as 5%, we conclude that the toothbrush data follow a multivariate normal distribution (see also the multivariate QQ-plot in the left-hand plot of Figure 1 in the Supplemental Material).

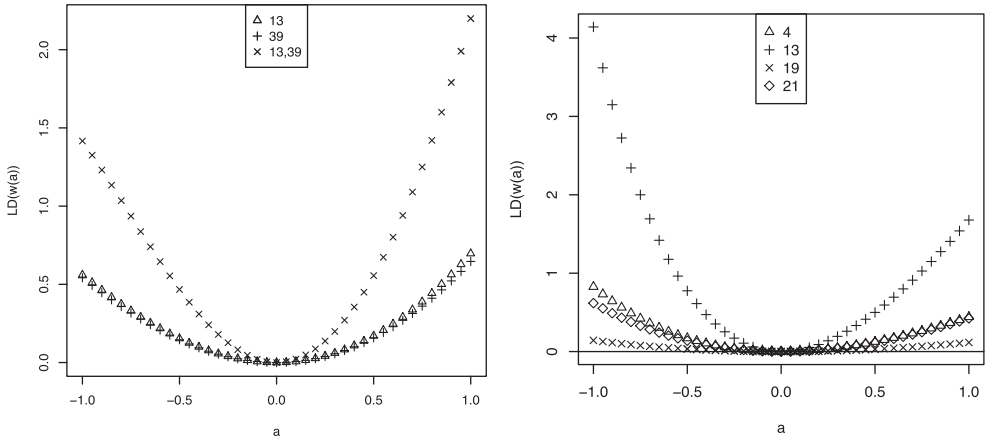

Figure 2. Toothbrush data: explanatory variable perturbation scheme (left-hand panel) and case weight perturbation scheme (right-hand panel). Plot of $LD(\boldsymbol{\omega}(a))$ versus a with $\boldsymbol{\omega}(a) = \boldsymbol{\omega}_0 + a\boldsymbol{l}$.

The second data set, which will be referred to as mouth rinse data, was presented in Ref. [20] and refers to a pretest/posttest study designed to compare three types of mouth rinses with respect to their efficacy in removing dental plaque. In that study, 105 adults were randomized to 2 experimental mouth rinses, A or B, or a control mouth rinse and the plaque indices were taken at the beginning of the study, after 3 months and after 6 months from the beginning of the study with the use of one of these mouth rinses. Thirty-six subjects used the control mouth rinse, while 33 (36) individuals used the experimental mouth rinse A (B). So, in this experiment, each subject used only one of the mouth rinses, but the data were taken longitudinally as the dental plaque indices were measured at the baseline, after three months and after six months from the baseline. For the same reasons that were discussed earlier, the null intercept measurement error model was proposed by Aoki *et al.* [2]. Also, as each subject was evaluated at baseline and two follow-up times, there is a possible dependence on the outcome measurements. So the structural model was considered. For the mouth rinse A (B), the p-value for the AD test was 0.608 (0.417) and for the CVM test was 0.582 (0.402) and for the control mouth rinse, these values were 0.140 (AD test) and 0.103 (CVM test). So, considering the significance level as 5%, we conclude that the data set referring to each of the mouth rinses follows a multivariate normal distribution (see also the multivariate QQ-plot in Figure 2 in the Supplemental Material).

Considering the mouth rinse data, let

$$\boldsymbol{z}_{i_j} = \begin{pmatrix} x_{i_j} \\ y_{1i_j} \\ y_{2i_j} \end{pmatrix}$$

denote the observed vector for the jth individual that was subject to the ith treatment, $j = 1, \ldots, n_i$, $i = 1, \ldots, p$. In the mouth rinse experiment, $p = 3$, with $i = 1$ representing the control mouth rinse and $i = 2$ ($i = 3$) the experimental mouth rinse A (B), $n_1 = 36$, $n_2 = 33$ and $n_3 = 36$. Let $\boldsymbol{x}_i = (x_{i_1}, \ldots, x_{i_{n_i}})^T$ denote the observed vector at baseline for subjects who used the ith mouth rinse, $\boldsymbol{y}_i = (\boldsymbol{y}_{1i}^T, \boldsymbol{y}_{2i}^T)^T$ the observed vector for subjects who used the ith mouth rinse, with $\boldsymbol{y}_{1i} = (y_{1i_1}, \ldots, y_{1i_{n_i}})^T$ representing the dental plaque indices after

three months from the baseline and $\boldsymbol{y}_{2i} = (y_{2i_1}, \ldots, y_{2i_{n_i}})^T$ the dental plaque indices after six months from the baseline, while $\boldsymbol{\xi}_i = (\xi_{i_1}, \ldots, \xi_{i_{n_i}})^T$ represents the unobserved true value of the dental plaque index before the use of the ith mouth rinse, $i = 1, 2, 3$. Then, the model can be written as [2]

$$\begin{aligned} \boldsymbol{x}_i &= \boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \\ \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{e}_i, \quad i = 1, \ldots, p, \end{aligned} \tag{3}$$

where

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{\xi}_i & \boldsymbol{0}_p \\ \boldsymbol{0}_p & \boldsymbol{\xi}_i \end{pmatrix},$$

$\boldsymbol{\beta}_i = (\beta_{1i}, \beta_{2i})^T$, $\boldsymbol{\delta}_i = (\delta_{i_1}, \ldots, \delta_{i_{n_i}})^T$, $\boldsymbol{e}_i = (\boldsymbol{e}_{1i}^T, \boldsymbol{e}_{2i}^T)^T$, with $\boldsymbol{0}_p$ denoting a vector composed by p 0's, $\boldsymbol{e}_{1i} = (e_{1i_1}, \ldots, e_{1i_{n_i}})^T$ and $\boldsymbol{e}_{2i} = (e_{2i_1}, \ldots, e_{2i_{n_i}})^T$, $\delta_{i_j} \overset{ind.}{\sim} N(0, \sigma_\delta^2)$, $e_{1i_j} \overset{ind.}{\sim} N(0, \sigma_{e_{1i}}^2)$, $e_{2i_j} \overset{ind.}{\sim} N(0, \sigma_{e_{2i}}^2)$, $\xi_{i_j} \overset{ind.}{\sim} N(\mu, \sigma_x^2)$, $\delta_{ij}, e_{1ij}, e_{2ij}$ and ξ_{ij} independent, $i = 1, \ldots, p$, $j = 1, \ldots, n_i$. So that the observed vector $\boldsymbol{z}_{i_j} \sim N_3(\boldsymbol{m}_i, \boldsymbol{V}_i)$, where $\boldsymbol{m}_i = \mu\boldsymbol{a}_i$ and $\boldsymbol{V}_i = \boldsymbol{A}_i + \sigma_x^2\boldsymbol{a}_i\boldsymbol{a}_i^T$ with $\boldsymbol{a}_i = (1, \boldsymbol{\beta}_i^T)^T$, $\boldsymbol{A}_i = \boldsymbol{D}(\sigma_\delta^2, \boldsymbol{\sigma}_{e_i}^{2T})^T$ with $\boldsymbol{\sigma}_{e_i}^2 = (\sigma_{e_{1i}}^2, \sigma_{e_{2i}}^2)^T$, $i = 1, \ldots, p$, $j = 1, \ldots, n_i$.

Then, the log likelihood function is given by

$$\ell(\boldsymbol{\theta}) = -\frac{3N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{p} n_i \log|\boldsymbol{V}_i| - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\boldsymbol{z}_{i_j} - \boldsymbol{m}_i)^T\boldsymbol{V}_i^{-1}(\boldsymbol{z}_{i_j} - \boldsymbol{m}_i) \tag{4}$$

with $\boldsymbol{\theta}_{(4p+3)\times 1} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T, \mu, \sigma_\delta^2, \sigma_x^2, \boldsymbol{\sigma}_{e_1}^{2T}, \ldots, \boldsymbol{\sigma}_{e_p}^{2T})^T$, $N = \sum_{i=1}^{p} n_i$, $|\boldsymbol{V}_i| = b_i|\boldsymbol{A}_i|$, $\boldsymbol{V}_i^{-1} = \boldsymbol{A}_i^{-1} - \sigma_x^2 b_i^{-1}\boldsymbol{B}_i$, where $b_i = 1 + \sigma_x^2\boldsymbol{a}_i^T\boldsymbol{A}_i^{-1}\mathbf{a}_i$ and $\boldsymbol{B}_i = \boldsymbol{A}_i^{-1}\boldsymbol{a}_i\boldsymbol{a}_i^T\boldsymbol{A}_i^{-1}$.

The third data set refers to a hydroscopic solid dosage.

Based on a risk map related to the product under analysis, a series of attributes are selected to be assessed during the stability study. The effects of variation in temperature, time, humidity, physical and chemical characteristics of the mixture and pH, among others, must be evaluated during the stability investigation process.

In a solid dosage product, several physical characteristics such as hardness, weight, thickness, disintegration, etc. can be evaluated during the stability analysis process. Among the attributes of stability studies for a hydroscopic product is the absorption of moisture over time. This study is important in order to choose the type of excipient mixture that presents lower absorption rate, as the moisture can interfere in the physical characteristics, in the analytical measurement of the content and in the dissolution behavior of the product, among others.

In this study, two mixtures of excipient are compared considering three follow-up times. The mixtures were kept in a laboratory environment, at 25°C and 55% relative humidity, during the 14 days of the experiment. The mass was weighted on an analytical balance in the beginning of the study, after 7 days and after 14 days with the product fully exposed to the condition of the laboratory environment. Furthermore, the balance induces a measurement error when measuring the mass of the oral solid drug product. Also, as the data were collected longitudinally, there is a correlation between measurements taken over time, but

not between different mixtures of excipient. So the model defined in (3) was considered. The p-value for the AD test was 0.737 (0.678) and for the CVM test was 0.888 (0.508) for the first mixture (solid dosage A) and the second mixture (solid dosage B), respectively. So, considering the significance level 5%, we conclude that the data set referring to each of the mixtures follows a multivariate normal distribution (see also the multivariate QQ-plot in Figure 1 in the Supplemental Material).

The proposed methodology was applied to these three data sets and a simulated data set to deal with masked influential observations in the multivariate measurement error model setting, though it is important to emphasize that the proposed methodology may be applied to any model where the local influence analysis is appropriate.

Section 2 gives a brief description of the two diagnostic measures, conformal normal curvature and forward search algorithm that will be used in the proposed methodology. Section 3 introduces the methodology. Applications with the real data sets just described and the simulated data set will be presented in Section 4. Finally, in Section 5, we discuss the obtained results.

2. Diagnostic analysis

In this section, we briefly describe the conformal normal curvature introduced by Poon and Poon [33] and the forward search algorithm proposed by Atkinson and Riani [5].

2.1. Conformal normal curvature

Let $LD(\boldsymbol{\omega}) = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\}$ denote the likelihood displacement function, where $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ and $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) = \log L(\boldsymbol{\theta}|\boldsymbol{\omega})$ with $L(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta}|\boldsymbol{\omega})$ representing the likelihood function and the perturbed likelihood function, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ the maximum likelihood estimates (MLEs) of $\boldsymbol{\theta}_{t\times 1}$ under the unperturbed and the perturbed models, respectively. $\boldsymbol{\omega} = (w_1, w_2, \ldots, w_r)^\top \in \Omega$ is the vector of perturbations, restricted to some open subset Ω of $\mathbb{R}^r$. A vector of no perturbation $\boldsymbol{\omega}_0$ is assumed, such that $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}_o) = \ell(\boldsymbol{\theta})$ and also $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$ is twice continuously differentiable. The graph of $\alpha(\boldsymbol{\omega}) = (\boldsymbol{\omega}^\top, LD(\boldsymbol{\omega}))^\top$, as $\boldsymbol{\omega}$ vary in Ω, is called the influence graph.

One way to investigate the local behavior of an influence graph around $\boldsymbol{\omega}_0$ is to select a direction $\boldsymbol{l}$ in Ω passing through $\boldsymbol{\omega}_0$.

Cook [16] showed that the normal curvature in the direction $\boldsymbol{l}$ can be written as

$$C_{\boldsymbol{l}} = 2|\boldsymbol{l}^T \boldsymbol{\Delta}^T \ddot{L}^{-1} \boldsymbol{\Delta} \boldsymbol{l}|,$$

where $-\ddot{L}$ is the observed information matrix, with $\ddot{L} = \partial^2 \ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, $\boldsymbol{\Delta} = \partial^2 \ell(\boldsymbol{\theta} \mid \boldsymbol{\omega})/\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^T|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0}$ and $\|\boldsymbol{l}\| = 1$.

Cook suggested the use of the maximum normal curvature, $C_{\max}$ (which is the maximum absolute eigenvalue of $\boldsymbol{\Delta}^T \ddot{L}^{-1} \boldsymbol{\Delta}$), and the associated eigenvector, $\boldsymbol{l}_{\max}$, to detect influential observations, as this is the direction that gives the greatest local change in the likelihood displacement.

Based on the work of Cook [16], Poon and Poon [33] proposed the use of the conformal normal curvature and they proved that the conformal normal curvature at a point $\boldsymbol{\omega}_0$ of

an influence graph in the direction $\boldsymbol{l}$ can be written as

$$B_{\boldsymbol{l}} = - \left. \frac{\boldsymbol{l}^T \boldsymbol{\Delta}^T \ddot{L}^{-1} \boldsymbol{\Delta} \boldsymbol{l}}{\sqrt{tr(\boldsymbol{\Delta}^T \ddot{L}^{-1} \boldsymbol{\Delta})^2}} \right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, \boldsymbol{\omega}=\boldsymbol{\omega}_0}.$$

Furthermore, the authors showed that $0 \leq | B_{\boldsymbol{l}} | \leq 1$ for any direction $\boldsymbol{l}$. Let $\lambda_h, h = 1, \ldots, r$, be the absolute value of the normalized eigenvalue of the matrix

$$\ddot{F} = \boldsymbol{\Delta}^T \ddot{L}^{-1} \boldsymbol{\Delta}, \tag{5}$$

such that

$$\lambda_{\max} = \lambda_1 \geq \cdots \geq \lambda_k \geq q/\sqrt{r} > \lambda_{k+1} \geq \cdots \geq \lambda_r \geq 0$$

and a_{hj} the jth element of the normalized eigenvector corresponding to λ_h. Poon and Poon [33] defined that an eigenvector $\boldsymbol{l}$ is q influential if $| B_{\boldsymbol{l}} | \geq q/\sqrt{r}$ and the aggregate contribution of the jth basic perturbation vector (column vector in $\mathbb{R}^r$ with the jth entry equal to 1 and all other entries equals to zero) to all q influential eigenvectors as

$$m(q)_j = \sqrt{\sum_{h=1}^{k} \lambda_h a_{hj}^2}. \tag{6}$$

If we allow q to be sufficiently large such that $\lambda_{\max} = \lambda_1 \geq q/\sqrt{r} \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$, then only the direction corresponding to the largest eigenvalue is considered in the analysis. Depending on the value of q, it is possible to consider the aggregate contribution of two largest eigenvalues and the associated eigenvectors, three largest eigenvalues and the associated eigenvectors and so on. If the value of q is small enough, the aggregate contribution of all the eigenvalues and the associated eigenvectors is considered.

We define $\boldsymbol{m}(q)^*$ as

$$\boldsymbol{m}(q)^* = (m(q)_1, \ldots, m(q)_r)^T. \tag{7}$$

So, if the contributions of all basic perturbation vectors are the same, each one would be equal to $\overline{m}(q) = \sqrt{\sum_{h=1}^{k} \lambda_h / r}$. Based on this result, Zhu and Lee [41] proposed some criteria to decide about potentially influential observations. In this paper, the benchmark will be considered as

$$\overline{m}(q) + 2sd \tag{8}$$

with sd denoting the standard deviation of the elements of the vector $\boldsymbol{m}(q)^*$.

2.2. Forward search

Atkinson and Riani [5] proposed the forward search algorithm to detect masked influential observations, considering regression models with the use of the least-squares methods and residuals. It starts by fitting the model to a small robustly chosen subset, supposedly free of outliers (basic set) and proceeds adding observations until all the observations are included.

Assume that we have a data set with n observations and let t denote the number of parameters. Afterwards, the procedure starts by obtaining all possible subsets of size t from n, $\binom{n}{t}$, or if this number is too big, a large number of subsets as 1000 is suggested. Subsequently, the least-squares estimate of the parameters for each subset of size $s = t$ is obtained. Moreover, for each subset, the residuals are calculated considering the parameter estimate obtained using that subset, but with all the n observations. Therefore, for each subset of size $s = t$, there will be a set of n residuals. The subset with least median square of the observational residuals is chosen to be the initial subset, the basic set.

This procedure is repeated until all the observations are included into the basic set ($t \leq s \leq n$), and the evolution of the quantities, such as parameter estimates, residuals and inferences are monitored as a function of the subset size. The procedure avoids the inclusion of outliers in the first steps, but the initial subsets does not affect the final steps where the most important information is concentrated. In the last step, we have a set with the whole observations and the estimation is obtained with the whole data set.

3. Conformal normal curvature with forward search

In Section 2, the conformal normal curvature introduced by Poon and Poon [33] and the forward search proposed by Atkinson and Riani [5] were briefly described. In this section, we introduce the proposed methodology, conformal normal curvature with forward search (CNCFS), to detect masked individually influential observations or groups of influential observations.

Many of the methodologies developed to obtain masked outliers divide the data set in a clean subset free of outliers, the basic set, and another data set composed by the remaining observations with potential outliers. A criterion is defined by which new observations are introduced into the basic set and it is incremented until all the observations are included. Therefore, in the last step, the parameter estimates are obtained and the analysis considering the whole data set is developed.

Considering the proposed methodology, first the perturbation scheme to be used is chosen. Zhu *et al.* [42] addressed the appropriate choice of a perturbation vector and used the first and second derivatives of the objective function to construct influence measures. They concluded that, for example, in the location-scale model, the case weight perturbation, the variance perturbation and the response variable perturbation are appropriate perturbations. Other models and other perturbations were also considered.

The case weight perturbation scheme is one of the most commonly used perturbation schemes. Another way to perturb the model is to consider heterogeneous variance by perturbing the variance terms. In addition, it is important that a small perturbation in the data set does not change the inference results, which leads to the perturbation in the explanatory variable and the response variable.

After defining the perturbation scheme to be used, the methodology starts at $s = s_0$ in the first step and it ends at $s = n$ in the last step. So that

$$s = s_0, s_0 + 1, \ldots, n - 1, n.$$

In each step, following Atkinson and Riani [5], either we sample all possible $\binom{n}{s}$ subsets of size s or if $\binom{n}{s} \geq 1000$, 1000 subsets are sampled. The size of the initial subset s_0 can be

chosen to be the number of parameters t or $t+1$, $t+2$ (see Ref. [5]). The choice of the initial subsets does not greatly influence the search and it does not affect the final steps where the most important findings of the analysis are concentrated. Let us denote the number of subsets in the sth iteration by b_s. Therefore, in the first step, there are b_{s_0} subsets of size s_0 and in the last iteration there are b_n subsets of size n ($b_n = 1$).

In the first step, the algorithm starts by sampling b_{s_0} samples of size s_0. Thereafter, the MLE of the parameters is obtained for each of the b_{s_0} subsets. In addition, for each of the b_{s_0} subsets, the value q is chosen, such that the aggregate contribution of the jth basic perturbation vector (6) would include the k largest eigenvalues. Furthermore, for each subset it is calculated the $\boldsymbol{m}(q)^*$ defined in (7) considering all the n observations (the whole data set), but with the MLE obtained for that subset and will be denoted by $\boldsymbol{m}(q)_c^*$, $c = 1, \ldots, b_{s_0}$. Subsequently, each of the b_{s_0} vectors $\boldsymbol{m}(q)_c^*$ are ordered and the least median vector is chosen, which will be denoted by $\boldsymbol{m}(q)_{s_0}$.

The forward search moves to the next iteration with $s = s_0 + 1$ and so forth, until all the observations are included. In the end, it is obtained

$$\boldsymbol{m}(q) = (\boldsymbol{m}(q)_{s_0}, \boldsymbol{m}(q)_{s_0+1}, \ldots, \boldsymbol{m}(q)_n)^T;$$

where $\boldsymbol{m}(q)_n$ is the aggregate contribution obtained by considering the whole data set, i.e. the usual aggregate contribution using the entire data set to obtain the MLE of the parameters.

Other quantities of interest can be obtained, such as

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_{s_0}, \ldots, \hat{\boldsymbol{\theta}}_n)^T,$$

where $\hat{\boldsymbol{\theta}}_s$, $s = s_0, \ldots, n$, represents the estimated parameters from the chosen set at each iteration. These quantities can be summarized using the forward plot.

The proposed methodology is summarized in the *Algorithm*.

In order to develop the CNCFS for the multivariate null intercept measurement error models defined in Section 1, four perturbation schemes were considered: explanatory variable perturbation scheme, case weight perturbation scheme, variance perturbation scheme and response variable perturbation scheme.

First, considering the model defined in (3) and (4), the necessary matrices to calculate the aggregate contribution of the jth basic perturbation vector, $j = 1, \ldots, r$, were obtained in closed form expressions.

The elements of the observed information matrix

$$-\ddot{L} = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \quad \text{with } \boldsymbol{\theta}_{(4p+3)\times 1} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T, \mu, \sigma_\delta^2, \sigma_x^2, \boldsymbol{\sigma}_{e_1}^{2T}, \ldots, \boldsymbol{\sigma}_{e_p}^{2T})^T$$

were calculated and are given in the Supplemental Material.

Moreover, let

$$\boldsymbol{\Delta}_{((4p+3)\times N)} = \left.\frac{\partial^2 \ell(\boldsymbol{\theta} \mid \boldsymbol{\omega})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}^T}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, \boldsymbol{\omega}=\boldsymbol{\omega}_0} = (\boldsymbol{\Delta}_{\theta 1_1}, \ldots, \boldsymbol{\Delta}_{\theta p_{n_p}})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, \boldsymbol{\omega}=\boldsymbol{\omega}_0},$$

Algorithm

$s = s_0$ (size of the initial subset)

- **Step 1**
  if $\binom{n}{s} \geq 1000$,
  $b_s = 1000$, else $b_s = \binom{n}{s}$.
- **Step 2**
  Sample b_s subsets of size s from the data set.
- **Step 3**
  For each of the b_s subsets sampled in Step 2:
  (a) obtain the MLEs of the parameters;
  (b) obtain the normalized eigenvalue and the corresponding normalized eigenvector of $\ddot{F}$ (5) using the whole data set (n observations) but the MLEs obtained in (a) using the subset;
  (c) according to the number of eigenvalues and the corresponding eigenvectors to be considered in the analysis, obtain the aggregate contribution according to (6) and name the (7) obtained in this iteration by $\boldsymbol{m}(q)_c^*$, where the index c corresponds to the subset ($c = 1, \ldots, b_s$).
  (d) order the vector $\boldsymbol{m}(q)_c^*$.
- **Step 4**:
  Considering the b_s vectors $\boldsymbol{m}(q)_c^*$, $c = 1, \ldots, b_s$ obtained in Step 3, choose the one with the least median vector and denote it by $\boldsymbol{m}(q)_s$ (in the first iteration $s = s_0$ and in the last iteration $s = n$).
- **Step 5**:
  Go to the Step 1 with $s = s_0 + 1$ until $s = n$.

In the end, we obtain $\boldsymbol{m}(q) = (\boldsymbol{m}(q)_{s_0}, \boldsymbol{m}(q)_{s_0+1}, \ldots, \boldsymbol{m}(q)_n)^T$. Plot $\boldsymbol{m}(q)$ vs $\boldsymbol{s}$, with $\boldsymbol{s} = (s_0, s_0 + 1, \ldots, n-1, n)^T$.

where $\boldsymbol{\Delta}_{\theta i_j} = (\boldsymbol{\Delta}^T_{\beta_1 i_j}, \ldots, \boldsymbol{\Delta}^T_{\beta_p i_j}, \boldsymbol{\Delta}_{\mu i_j}, \boldsymbol{\Delta}_{\sigma_x^2 i_j}, \boldsymbol{\Delta}_{\sigma_\delta^2 i_j}, \boldsymbol{\Delta}^T_{\sigma_{e_1}^2 i_j}, \ldots, \boldsymbol{\Delta}^T_{\sigma_{e_p}^2 i_j})^T_{((4p+3)\times 1)}$, $i = 1, \ldots, p$, $j = 1, \ldots, n_i$, and $N = \sum_{i=1}^{p} n_i$.

After algebraic manipulations, the elements of the matrix $\boldsymbol{\Delta}$ for each perturbation scheme were obtained and are given in the Supplemental Material.

Next, to show the usefulness of the proposed methodology, we apply the CNCFS to the three real data sets described in Section 1 and a simulated data set. The routines were implemented in R Core Team [34].

4. Application

In this section, the proposed methodology is illustrated with three real applications and the simulated data. The first data set to be examined is the toothbrush data as it is simpler to analyze.

4.1. Toothbrush data

Aoki *et al.* [1] proposed the use of the measurement error model with null intercept to compare the efficacy of two types of toothbrushes, the experimental toothbrush and the regular toothbrush, described in Section 1. Considering the model defined in (1) and (2), the CNCFS methodology was applied to the perturbation schemes defined in Section 3.

We considered two extremes in this methodology. The total contribution, i.e. we assumed that $q = 0$ so that all the eigenvalues and the associated eigenvectors are included in the analysis and then we allowed q to be sufficiently large so that only the contribution of the largest eigenvalue and the associated eigenvector are considered, i.e. the direction which causes the greatest local change in the likelihood displacement.

First, considering the explanatory variable perturbation scheme, the forward plot of CNCFS and the index plot of $\boldsymbol{m}(q)$ were obtained and are shown in Figure 1.

The forward plot of CNCFS (top panels of Figure 1) starts with a subset of size $s = 10$ and in each step of the algorithm, the sample size is incremented by 1, so that in the last step the analysis and the estimation of the parameters are done with the whole sample of size $s = 26$, which means that the points in the last iteration of the forward plot of CNCFS are the same as the points in the usual index plot of the aggregate contribution (bottom panels of Figure 1). In the index plots of $\boldsymbol{m}(q)$ (bottom panels of Figure 1), the indices 1–26 in the x axis refer to the data associated to the experimental toothbrush, while the indices 27–52 refer to the data corresponding to the conventional toothbrush for the same individuals in the same order.

Considering the left-hand panel (top), which refers to the forward plot of CNCFS with q sufficiently large so that only the contribution of the eigenvector associated with the largest eigenvalue is included, clearly observation 39 is above the benchmark for almost the entire CNCFS evolution, however it is masked in the penultimate iteration and the observation 13 pops up.

According to the index plot of $\boldsymbol{m}(q)$ (left-hand bottom panel), which refers to the last iteration of the forward plot of CNCFS and also is the usual index plot of the aggregate contribution, the conclusion is that only the observation 13 is above the benchmark. But with the use of the forward plot of CNCFS, it is possible to see that observation 39 was masked in the final steps.

Next, the right-hand panels of Figure 1 show the forward plot of CNCFS and the index plot of $\boldsymbol{m}(q)$ when $q = 0$, i.e. when all eigenvalues and the associated eigenvectors are included in the analysis. In this case, observation 13 is above the benchmark from $s = 16$ to the end of CNCFS evolution, while observation 39 is mostly under the benchmark from $s = 16$ to the end of the evolution, though very close to the benchmark.

The corresponding index plot of $\boldsymbol{m}(q)$ (right-hand bottom panel) shows that observation 13 is above the benchmark. However, with this plot, it is not possible to have an overview of the evolution of the influence of this observation as the number of elements in the subset increases.

Individual 13 corresponds to the preschooler with the second largest plaque index before the use of the experimental toothbrush and the largest post-toothbrushing dental plaque index among the individuals who used the experimental toothbrush. Observation 39 also corresponds to the individual 13 but with the use of the conventional toothbrush and in this case he is the child with the smallest pre-toothbrushing dental plaque index among

the individuals who used the conventional treatment and also the only child that had no reduction under the conventional treatment.

To analyze the influence of these observations in the likelihood displacement, the left-hand panel of Figure 2 shows the plot of $LD(\boldsymbol{\omega}_0 + a\boldsymbol{l})$ versus $a \in [-1, 1]$ along the directions $\boldsymbol{l} = \boldsymbol{l}_k$, with $k = 13$, 39 and $(13, 39)$, with $\boldsymbol{l}_k$ denoting a null vector of size 26 with the kth element(s) replaced by 1. It can be seen that the influence of observations 13 and 39 are very close. Moreover, if we perturb together these two observations, they become much more influential.

Next, we show the forward plot of CNCFS and the index plot of $\boldsymbol{m}(q)$ for the contribution of the eigenvector associated with the largest eigenvalue, where no observation was masked during the CNCFS evolution.

In the left-hand panels of Figure 3, the response variable perturbation scheme was considered. In the forward plot of CNCFS (left-hand top panel), observation 44 are above the benchmark during the entire CNCFS evolution. Also, there was no masked influential observations. Considering the index plot of $\boldsymbol{m}(q)$ (left-hand bottom panel), the observation 44 is above the benchmark. Moreover, clearly small local changes in the response variables associated with the conventional toothbrush have a larger effect on the parameter estimates. Observation 44 corresponds to the preschooler 18 who had the greatest value of the dental plaque index after the use of the conventional toothbrush.

The right-hand panels of Figure 3 show the forward plot of CNCFS (right-hand top panel) and the index plot of $\boldsymbol{m}(q)$ (right-hand bottom panel) where the indices 1–26 in the x axis are the index of the observations considering the case weight perturbation scheme. Observation 13 is above the benchmark during the entire CNCFS evolution. Both of the plots (top and bottom) clearly show that observation 13 is influential. Considering the forward plot os CNCFS observation 4 is detached from the rest of the observations, though below the benchmark. The same happens with observations 21 and 19, but possibly less influential.

Figure 2 (right-hand panel) shows the plot of the likelihood displacement along the directions $\boldsymbol{l} = \boldsymbol{l}_k$, with $k = 13$, 4, 21 and 19. Notice that the value of the likelihood displacement increases quickly, when the perturbation is made in the direction of observation 13 compared to the other observations. Comparing the rest of the observations, observation 4 is the most influential among the observations 4, 21 and 19, and it is followed by observations 21 and 19.

4.2. Simulated data

Next, we generated a simulated data set considering the model defined in (1) and (2) for the toothbrush data according to Figure 4, with 26 observations. Considering variance perturbation scheme and the contribution of the eigenvector associated with the largest eigenvalue, two largest eigenvalues, three largest eigenvalues and all the eigenvalues, the forward plot of CNCFS was obtained in Figure 5. The top left panel shows the forward plot of CNCFS for the contribution of the eigenvector associated with the largest eigenvalue, the top right-hand panel shows the forward plot of CNCFS for aggregate contribution of the two largest eigenvalues, the bottom left-hand panel shows the forward plot of CNCFS for the aggregate contribution of the three largest eigenvalues and the right-hand bottom panel shows the forward plot of CNCFS for the total contribution ($q = 0$).

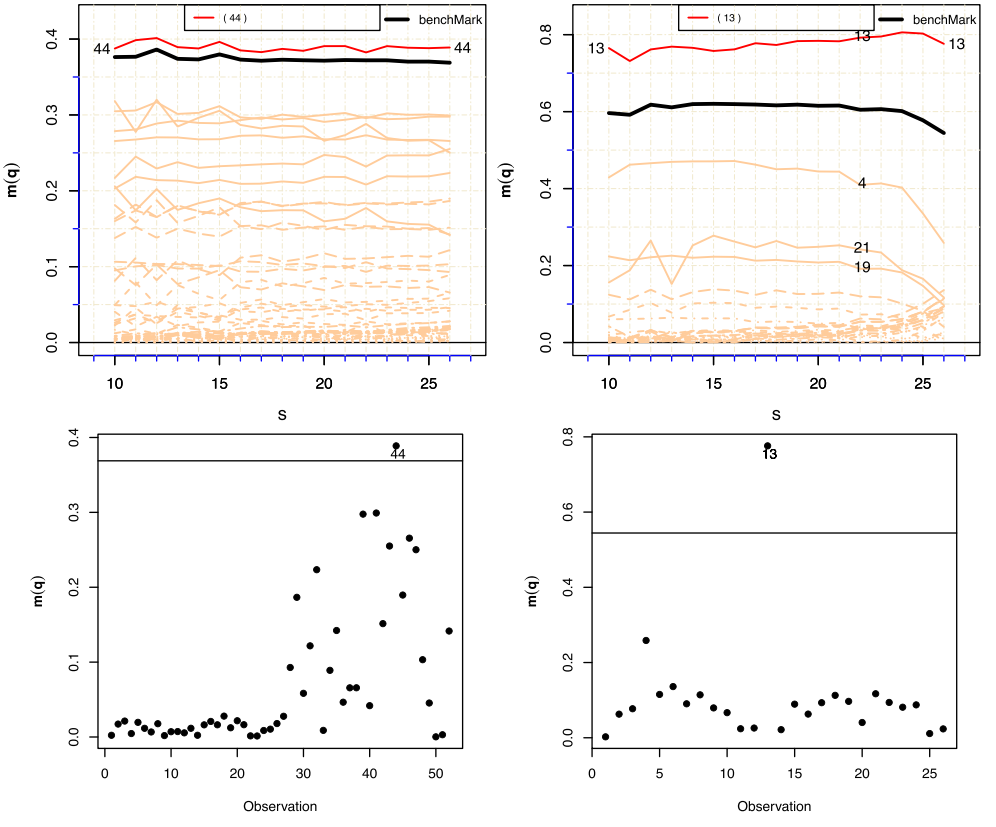

Figure 3. Toothbrush data: response variable perturbation scheme (left-hand panels) and case weight perturbation scheme (right-hand panels). CNCFS forward plot (top) and index plot of $\boldsymbol{m}(q)$ (bottom) for the contribution of the eigenvector associated with the largest eigenvalue.

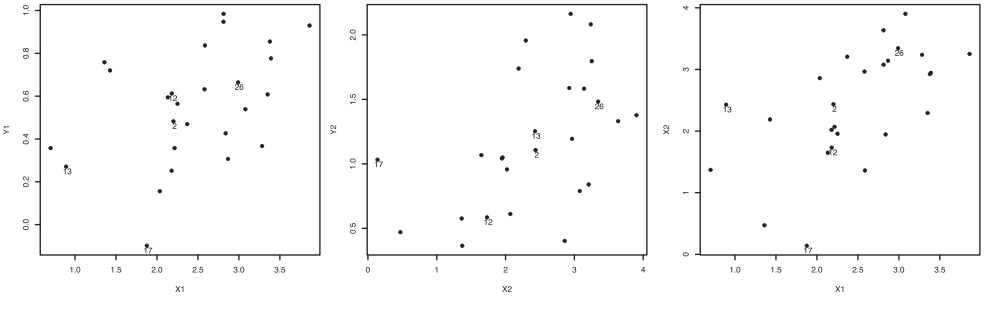

Figure 4. Scatter plot of the simulated data.

Notice that in all cases, although below the bench mark, observations 13 and 17 are depicted from the rest of the observations.

So it was considered two scenarios to deliberately change the value of the observations.

In the first scenario, we modified observations 13 and 17, while in the second scenario we considered three randomly chosen observations, 2, 12 and 26, to deliberately change. These observations are depicted in Figures 4, 5, 6 and 7. Figure 6 shows the scatter plot

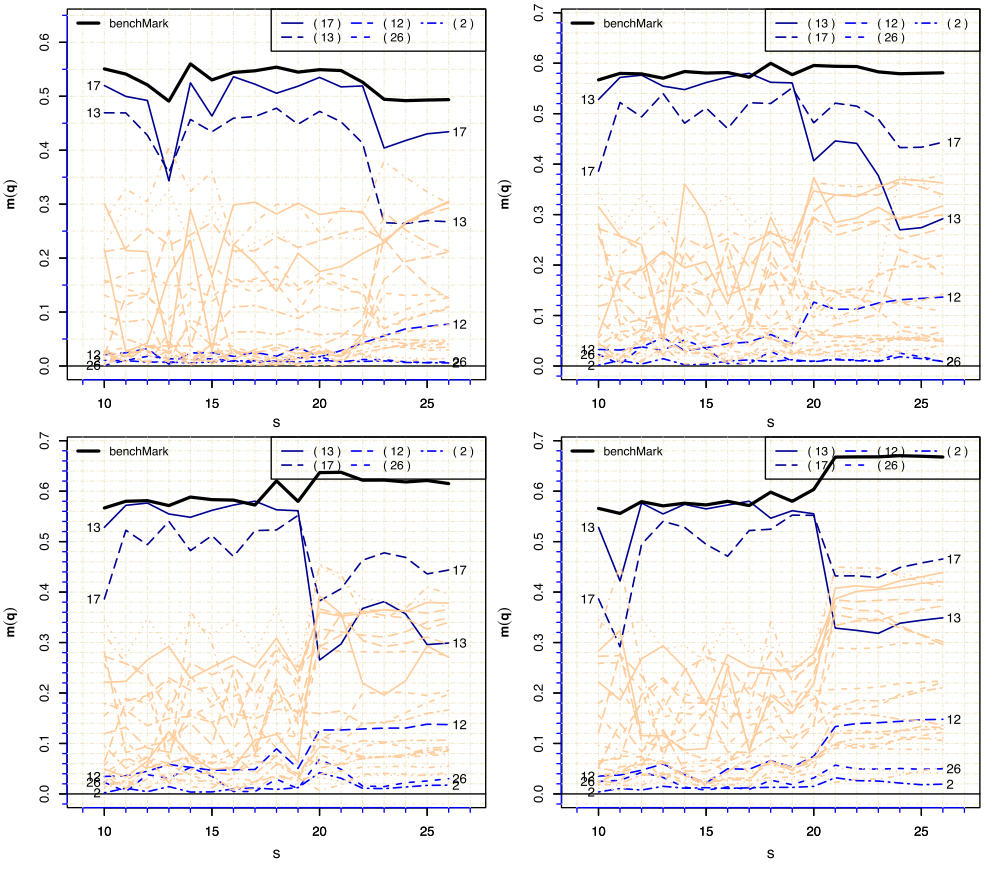

Figure 5. Simulated data: variance perturbation scheme. CNCFS forward plot for the contribution of the eigenvector associated with the largest eigenvalue (top left), aggregate contribution of the two largest eigenvalues and the associated eigenvectors (top right), aggregate contribution of the three largest eigenvalues and the associated eigenvectors (bottom left) and the total contribution ($q = 0$) in the bottom right-hand panel.

of the observations after altering observations 13 and 17, whereas Figure 8 presents the forward plot of CNCFS for the contribution of the eigenvector associated with the largest eigenvalue (top left-hand panel), the two largest eigenvalues (top right-hand panel), the three largest eigenvalues (bottom left-hand panel) and the total contribution ($q = 0$, right-hand bottom panel).

First scenario:

Clearly, observation 13 appears as possibly influential observation in all the panels and observation 17 is masked in the bottom panels, while in the top panels it is below the benchmark although detached from the rest of the observations.

Figure 8 shows the plot of the likelihood displacement along the directions $\boldsymbol{l} = \boldsymbol{l}_k$, with $k = 13$ and 17 for the original data set (left-hand panel) and after changing the observations 13 and 17 (right-hand panel).

Notice that the value of the likelihood displacement increases quickly after changing observations 13 and 17. In the left-hand panel, the highest value of $LD(\boldsymbol{\omega}(a))$ at $a = -1$

[!t]

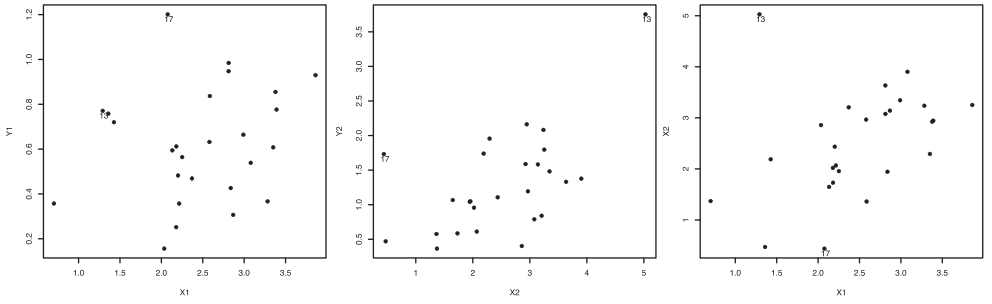

Figure 6. Scatter plot of the simulated data, where observations 13 and 17 were deliberately changed.

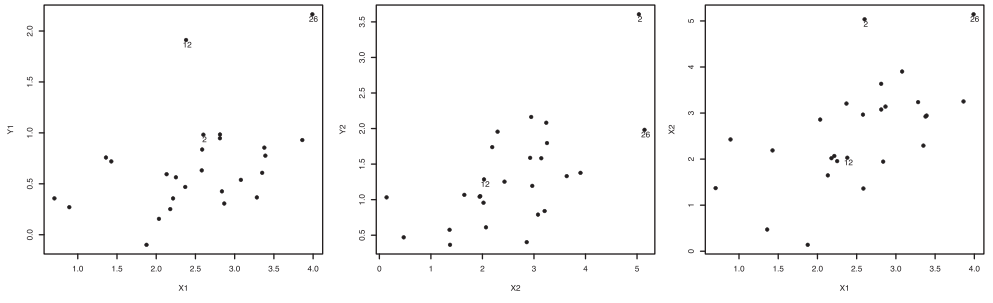

Figure 7. Simulated data: variance perturbation scheme (left-hand panel) and variance perturbation scheme after changing observations 13 and 17 (right-hand panel). Plot of $LD(\boldsymbol{\omega}(a))$ versus a with $\boldsymbol{\omega}(a) = \boldsymbol{\omega}_0 + a\boldsymbol{l}$.

for observation 17 is near 1.4 for the original data set, while for the data set after changing observations 13 and 17, it is near 19 for observation 13.

Next, we consider the second scenario where three randomly chosen observations, 2, 12 and 26, were deliberately changed.

Second scenario:

In the second scenario, three randomly chosen observations, 2, 12 and 26, which can be seen in Figures 4 and 5 were changed, giving rise to Figure 7.

The corresponding forward plot of CNCFS for the variance perturbation scheme was obtained for the contribution of the eigenvector associated with the largest eigenvalue (top left-hand panel), the two largest eigenvalues (top right-hand panel), the three largest eigenvalues (bottom left-hand panel) and for the total contribution ($q = 0$) in the right-hand bottom panel of Figure 10.

Comparing Figures 5 and 10, observation 2 is on the bottom of the forward plot of CNCFS during the entire CNCFS evolution algorithm for the original data set in all of the four considered directions (Figure 5), while after deliberately changing the observations 2, 12 and 26 it becomes a possibly influential observation in all of the four considered directions in Figure 10.

Moreover, to see how the influence of this observation changed after modifying the three observations, the likelihood displacement of the three altered observations were obtained

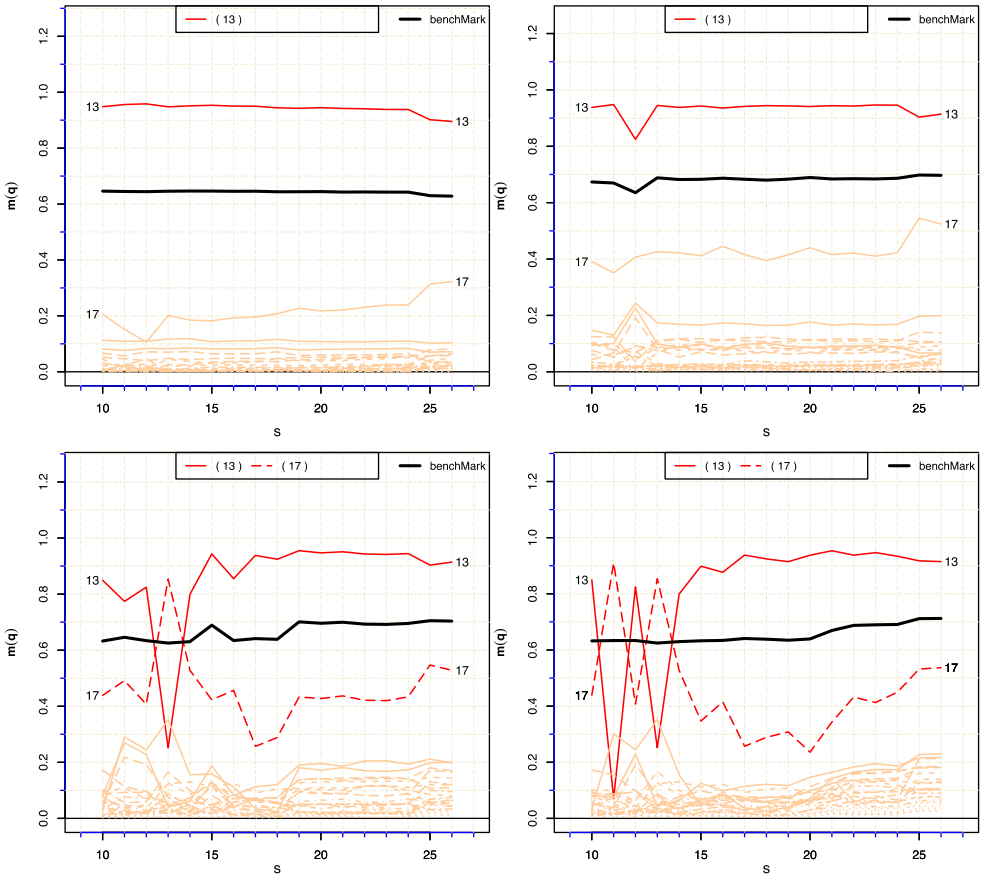

Figure 8. Simulated data after changing observations 13 and 17: variance perturbation scheme. CNCFS forward plot for the contribution of the eigenvector associated with the largest eigenvalue (top left), aggregate contribution of the two largest eigenvalues and the associated eigenvectors (top right), aggregate contribution of the three largest eigenvalues and the associated eigenvectors (bottom left) and the total contribution ($q = 0$) in the bottom right-hand panel.

before and after changing their values and they are shown in Figure 11. Observation 2 is the least influential observation among the three observations considering the original data set (left-hand panel of Figure 11), while after changing the value of the three observations, it became the most influential observation among these observations (right-hand panel of Figure 11). In addition, the highest value of $LD(\boldsymbol{\omega}(a))$ at $a = -1$ is near 0.03 for observation 2 considering the original data set, whereas for the modified data set, it is near 8 for observation 2.

According to Figure 5, among observations 2, 12 and 26, observation 12 seems to be the most influential, followed by the observations 26 and 2, which is confirmed on the left-hand panel of Figure 11. After changing these observations, Figure 10 shows that observation 2 may be influential, observation 12 was masked and observation 26 is below the benchmark during the whole CNCFS evolution, though detached from the rest of the observations. See also the right-hand panel of Figure 11.

The third data set to be analyzed is the mouth rinse data set.

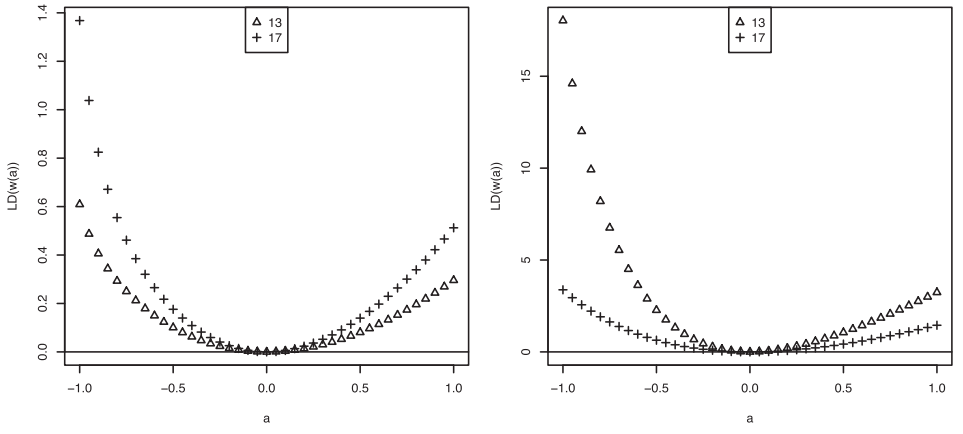

Figure 9. Scatter plot of the simulated data, where observations 2, 12 and 26 were deliberately changed.

4.3. Mouth rinse data

Considering the mouth rinse data set described in Section 1, the CNCFS methodology was applied to the four perturbation schemes described in Section 3.

In this case, we used the aggregate contribution of the two largest eigenvalues and the associated eigenvectors and also q sufficiently large so that only the contribution of the largest eigenvalue and its associated eigenvector are considered. The forward search started with a subset of size $s = 15$ and in each step of the algorithm the sample size was incremented by 1 until all the $N = 105$ observations were included in the last iteration, when $s = 105$. As a consequence, the last iteration gives the usual index plot of $\boldsymbol{m}(q)$ considering the whole data set. The MLE of the parameters was obtained using the EM algorithm described in Ref. [37].

Figure 12 shows the forward plot of aggregate contribution of the two largest eigenvalues and the associated eigenvectors considering the explanatory variables perturbation scheme. Clearly, observation 80 must be influential as it is mostly above the benchmark from iteration $s = 39$ until the penultimate iteration when it becomes masked.

However, the usual index plot of $\boldsymbol{m}(q)$ (Figure 13), which is the last iteration of the forward plot, would lead to the conclusion that none of the observations are influential, as all the observations are under the benchmark. The indices 1–36 in the x axis refer to the data associated with the individuals who used the control mouth rinse while the indices 37–69 (70–105) refer to the observations corresponding to the subjects who used the experimental mouth rinse A (B). Observation 80 belongs to the group who used the experimental mouth rinse B and among these individuals, he had the second highest value of the plaque index in the beginning of the treatment and also the smallest reduction in the dental plaque index from the beginning of the study to the end of the study.

Table 1 in the Supplemental Material shows that with significance level 10%, the removal of observation 80 changes the conclusion of one of the hypotheses of interest (see Ref. [37] for more details), i.e. the hypothesis that the dental plaque index reduction rate after six months using the experimental mouth rinse A and B are the same is no longer rejected. On

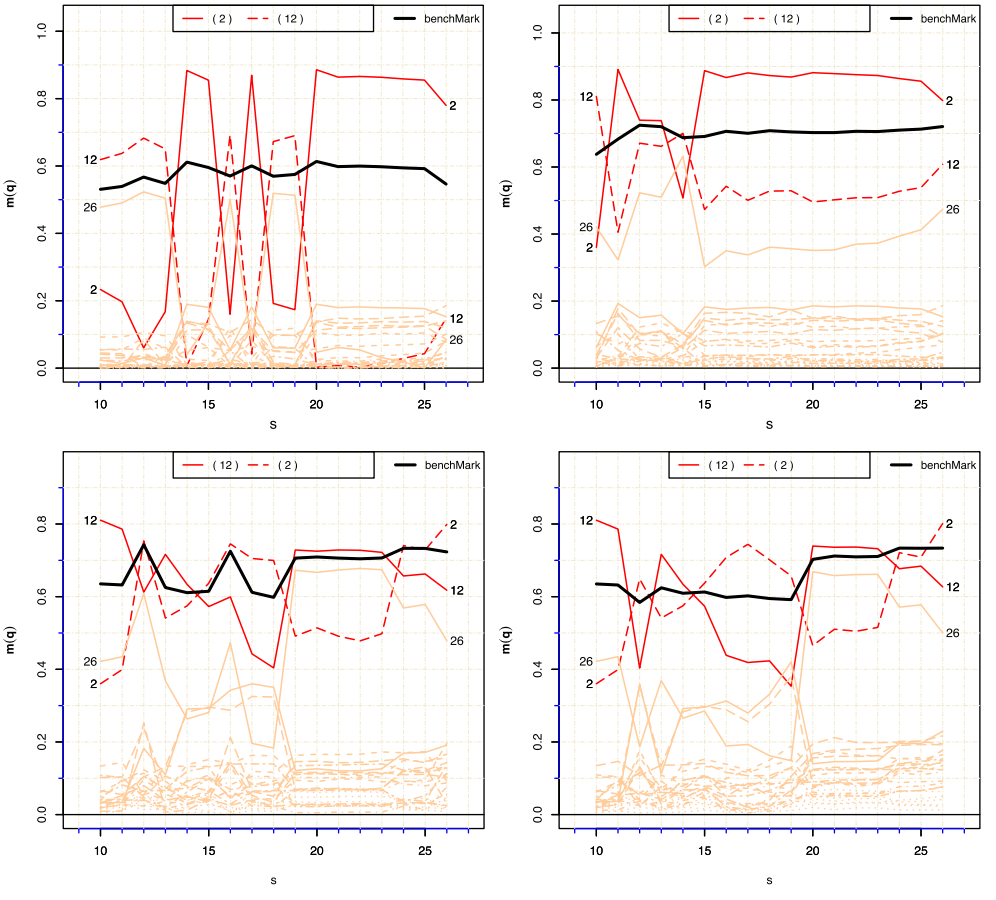

Figure 10. Simulated data after changing observations 2, 12 and 26: variance perturbation scheme. CNCFS forward plot for the contribution of the eigenvector associated with the largest eigenvalue (top left), aggregate contribution of the two largest eigenvalues and the associated eigenvectors (top right), aggregate contribution of the three largest eigenvalues and the associated eigenvectors (bottom left) and the total contribution ($q = 0$) in the bottom right-hand panel.

the other hand, the removal of this observation (global influence) gives a little change in the parameter estimates (not shown here). As can be seen, the forward plot of CNCFS is crucial to detect these kinds of observations that cannot be detected with the usual analysis.

Furthermore, in the Supplemental Material, we present the variance perturbation scheme, response variable perturbation scheme and case weight perturbation scheme, analyzing the influence of observations that were masked and observations that appear above the benchmark.

Depending on the situation, we can have many observations above the benchmark. In this case, one way to have an overview of the influence of each of the observations in each iteration is the heatmap. In the Supplemental Material, we show the heat map relative to the case weight perturbation scheme. It shows the degree of influence of each observation in each iteration. See the Supplemental Material for more detail.

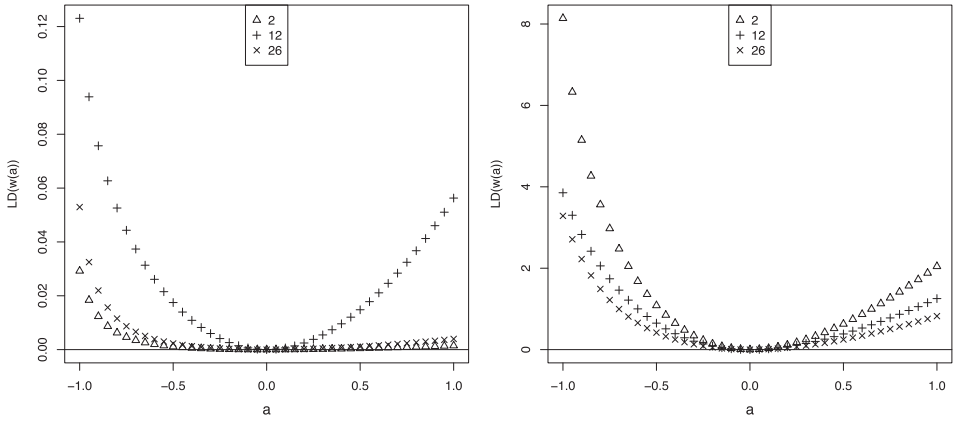

Figure 11. Simulated data: variance perturbation scheme (left-hand panel) and variance perturbation scheme after changing observations 2, 12 and 26 (right-hand panel). Plot of $LD(\boldsymbol{\omega}(a))$ versus a with $\boldsymbol{\omega}(a) = \boldsymbol{\omega}_0 + a\boldsymbol{l}$.

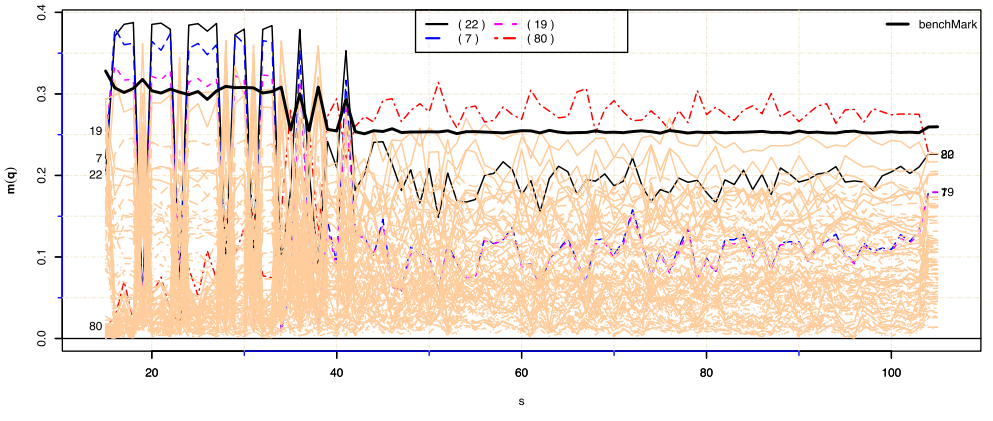

Figure 12. Mouth rinse data: CNCFS forward plot for explanatory variable perturbation scheme. Aggregate contribution of the two largest eigenvalues and the associated eigenvectors.

4.4. Hygroscopic solid dosage data

Finally, the last data set is from a stability study of a hygroscopic solid dosage, where the interest was to compare two mixtures of the excipient considering three follow-up times. The mass was weighted in the beginning of the study, after 7 days and after 14 days with the product fully exposed to the condition of the laboratory environment as described in Section 1. The model defined in (3) and (4) and the aggregate contribution of the three largest eigenvalues and the associated eigenvectors were considered.

The data set can be found in Online Resource.

First, considering the explanatory variable perturbation scheme, the forward plot of CNCFS was obtained and is shown in Figure 14. The forward search started with a subset of size $s = 11$ and in each step of the algorithm, the sample size was incremented by 1 until all the $N = 200$ observations were included in the last iteration when $s = 200$, i.e. in the last step, the analysis is done with the whole data set.

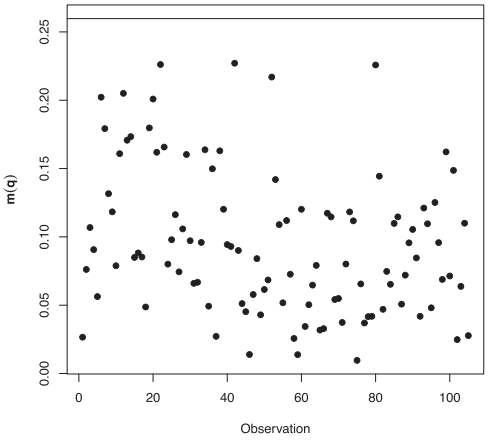

Figure 13. Mouth rinse data: index plot of $\boldsymbol{m}(q)$. Aggregate contribution of the two largest eigenvalues and the associated eigenvectors, explanatory variable perturbation scheme.

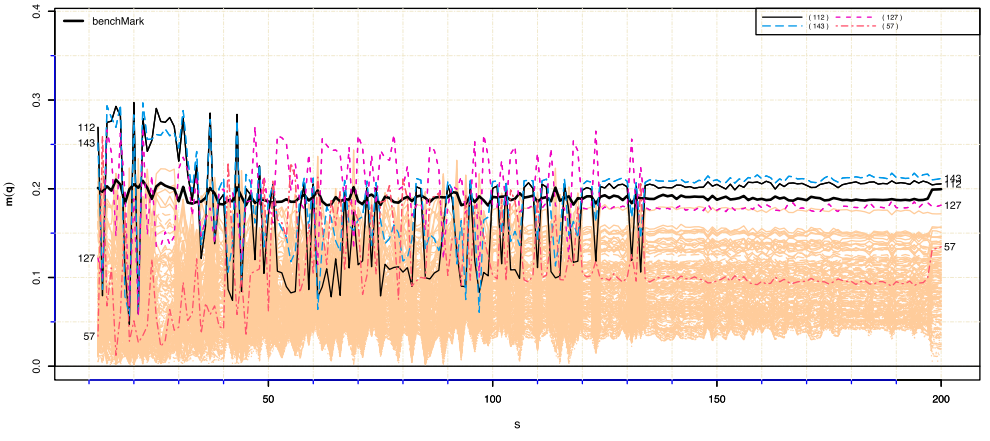

Figure 14. Hygroscopic solid dosage data: CNCFS forward plot for explanatory variable perturbation scheme. Aggregate contribution of the three largest eigenvalues and the associated eigenvectors.

Observation 127 is above the benchmark in most iterations from the beginning until iteration $s = 133$ of CNCFS evolution and then it is masked, while observations 112 and 143 are above the benchmark from iteration $s = 134$ until the end of CNCFS evolution. Observations 1–100 refer to the solid dosage with mixture of excipient A, while observations 101–200 refer to the solid dosage with mixture of excipient B. From now on they will simply be referred to as solid dosages A and B.

Observation 127 along with observation 112 are the only solid dosages whose weights decreased from the beginning of the study to after 7 days. Observation 57 also appears as influential in the middle part of CNCFS evolution, but with less intensity. Observation 57 is the solid dosage A with the highest weight at the beginning of the study and also after 7 and 14 days from the beginning of the study. In addition, it is the observation that absorbed the least moisture from the environment from the beginning of the study to the end of the study among the solid dosage A.

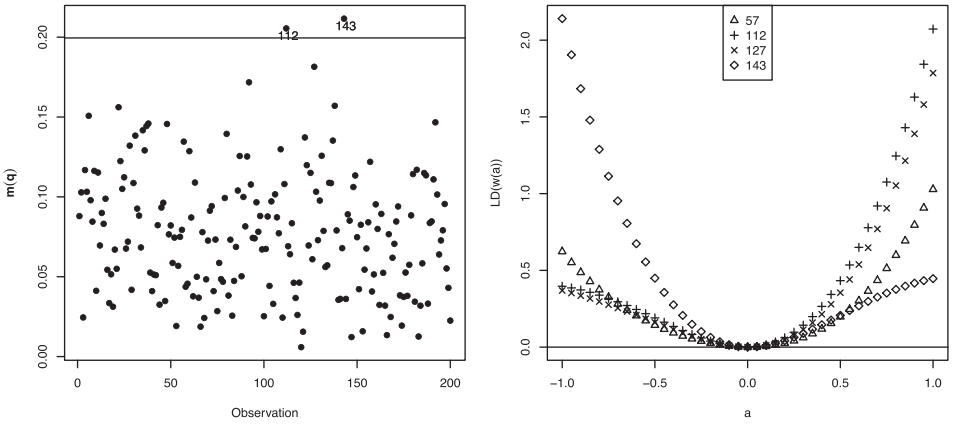

Figure 15. Hygroscopic solid dosage data: explanatory variable perturbation scheme. Index plot of $\boldsymbol{m}(q)$ for the aggregate contribution of the three largest eigenvalues and the associated eigenvectors (left-hand panel) and plot of $LD(\boldsymbol{\omega}(a))$ versus a with $\boldsymbol{\omega}(a) = \boldsymbol{\omega}_o + a\boldsymbol{l}$ (right-hand panel).

To see the influence of perturbing these observations in the likelihood displacement function, the plot of $LD(\boldsymbol{\omega}_0 + a\boldsymbol{l})$ versus $a \in [-1, 1]$ along the directions $\boldsymbol{l} = \boldsymbol{l}_k$, with $k = 57, 112, 127$ and 143, was obtained and it is shown in the right-hand panel of Figure 15. For positive values of a, observation 112 and 127 (which was masked) appear as the most influential ones among observations 57, 112, 127 and 143, followed by observations 57 that was masked. In the direction of negative values of a, observation 143 appears as the most influential observation among the considered observations. The usual index plot of $\boldsymbol{m}(q)$, left-hand panel of Figure 15, which refers to the last iteration of CNCFS, only shows that observations 112 and 143 may be influential as they are above the benchmark.

The variance perturbation scheme and the case weight perturbation scheme were also considered. The analysis can be found in the Supplemental Material.

5. Conclusion

It is well known that the influence analysis is an important step in the data analysis. As commented in Section 1, the analyzes of usual linear regression and linear measurement error models can be drastically altered by the presence of influential/outlier observations. The usual methodologies used to identify influential observations may be affected by the own observations that should be detected and fail to detect these influential observations. Also, if there is a group of observations that are jointly influential, but not individually influential, the global influence where a case is deleted one at a time will fail to detect these observations. In order to detect masked observations, Atkinson and Riani [5] proposed the use of the forward search in regression models based on the least-squares estimates and the associated residuals. On the other hand, Poon and Poon [33] introduced the use of conformal normal curvature with a benchmark to assess the local influence of minor perturbations in the data set or in the model. This methodology can identify groups of observations that may be jointly influential. Nevertheless, the local influence analysis can fail to detect masked influential observations as can be clearly seen in Section 4 and in the Supplemental Material.

In this paper, to overcome these shortcomings, we proposed the CNCFS methodology based on the forward search algorithm and the conformal normal curvature. An interesting point in the methodology is that even when there are no masked observations, it is possible to have an overview of the evolution of each observation in each step of the CNCFS (see Figure 3 (top)), giving an extra information, and also if there are masked influential observations, it is possible to see in which iterations the observations were above the bench mark (see Figures 12 and 3 of Supplemental Material, for instance).

In Section 4, where the proposed methodology were applied to real data sets, many masked observations were successfully detected, of which many of them changed the results of the inference as discussed in the Application Section and in the Supplemental Material. The forward plot of CNCFS was crucial to detect observations that could not be detected with the usual local influence analysis.

Hence, the contribution of this paper is to provide a graphical methodology based on easy to interpret plots to detect observations that are individually influential or a group of observations that are jointly influential.

The methodology were considered in measurement error models, however it is important to observe that it can easily be used in any statistical model where the local influence analysis [16] can be performed.

Acknowledgments

The authors would like to thank the Associate Editor and three referees for their constructive comments and recommendations that definitely helped improve the quality of the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

References

[1] R. Aoki, H. Bolfarine, and J.M. Singer, *Null intercept measurement error regression models*, Test. 10 (2001), pp. 441–457.
[2] R. Aoki, H. Bolfarine, J.A. Achcar, and L.P. Dorival Jr, *Bayesian analysis of a multivariate null intercept errors-in-variables regression model*, J. Biopharm. Stat. 13 (2003), pp. 767–775.
[3] R. Aoki, J. da Motta Singer, and H. Bolfarine, *Local influence for measurement error regression models for the analysis of pretest/posttest data*, J. Appl. Stat. Sci. 15 (2007), pp. 317–330.
[4] R. Aoki, J.P.M. Bustamante, and G.A. Paula, *Local influence diagnostics with forward search in regression analysis*, Stat. Pap. 63 (2022), pp. 1477–1497.
[5] A. Atkinson and M. Riani, *Robust Diagnostic Regression Analysis*, Springer, New York, 2000.
[6] A.C. Atkinson, M. Riani, and A. Cerioli, *Exploring Multivariate Data with the Forward Search*, Springer Science & Business Media, London, 2013.
[7] A.C. Atkinson, *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Clarendon Press Oxford, Oxford, 1985.
[8] T. Bellini, *Detecting atypical observations in financial data: the forward search for elliptical copulas*, Adv. Data Anal. Classif. 4 (2010), pp. 287–299.

[9] D.A. Belsley, E. Kuh, and R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, New York, 1980. http://onlinelibrary.wiley.com/book/
[10] R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. Monographs on Statistics and Applied Probability, Chapman and Hall, Boca Raton, FL, 2006. https://cds.cern.ch/record/1012043
[11] R.J. Carroll, A. Delaigle, and P. Hall, *Nonparametric prediction in measurement error models*, J. Am. Stat. Assoc. 104 (2009), pp. 993–1003.
[12] A. Cerioli, A. Farcomeni, and M. Riani, *Strong consistency and robustness of the forward search estimator of multivariate location and scatter*, J. Multivar. Anal. 126 (2014), pp. 167–183.
[13] S. Chatterjee and A.S. Hadi, *Influential observations, high leverage points, and outliers in linear regression*, Stat. Sci. 1 (1986), pp. 379–393.
[14] C.-L. Cheng and J.W. Van Ness, *Statistical Regression with Measurement Error*, John Wiley & Sons, Oxford, 1999.
[15] R.D. Cook, *Detection of influential observation in linear regression*, Technometrics. 19 (1977), pp. 15–18.
[16] R.D. Cook, *Assessment of local influence*, J. R. Stat. Soc. Ser. B Methodol. 48 (1986), pp. 133–169.
[17] R.D. Cook and S. Weisberg, *Residuals and influence in regression*, Monogr. Stat. Appl. Probab. New York, 1982.
[18] M. Galea-Rojas, H. Bolfarine, and M. de Castro, *Local influence in comparative calibration models*, Biom. J. 44 (2002), pp. 59–81.
[19] A. Grané, G. Manzi, and S. Salini, *Smart visualization of mixed data*, Stats. 4 (2021), pp. 472–485. https://www.mdpi.com/2571-905X/4/2/29
[20] A. Hadgu and G. Koch, *Application of generalized estimating equations to a dental randomized clinical trial*, J. Biopharm. Stat. 9 (1999), pp. 161–178.
[21] Y. Hu and T. Wansbeek, *Measurement error models: editors' introduction*, J. Econ. 200 (2017), pp. 151–153. http://www.sciencedirect.com/science/article/pii/S0304407617300817
[22] S. Johansen and B. Nielsen, *Asymptotic theory of outlier detection algorithms for linear time series regression models*, Scand. J. Stat. 43 (2016), pp. 321–348.
[23] G. Kelly, *The influence function in the errors in variables problem*, Ann. Stat. 12 (1984), pp. 87–100.
[24] J.A. Koziol, *A class of invariant procedures for assessing multivariate normality*, Biometrika. 69 (1982), pp. 423–427.
[25] F.V. Labra, R. Aoki, and H. Bolfarine, *Local influence in null intercept measurement error regression under a student_t model*, J. Appl. Stat. 32 (2005), pp. 723–740. DOI:10.1080/02664760500079639
[26] F.V. Labra, R. Aoki, V. Garibay, and V.H. Lachos, *Skew-normal distribution in the multivariate null intercept measurement error model*, Braz. J. Probab. Stat. 25 (2011), pp. 145–170.
[27] A.H. Lee and Y. Zhao, *Assessing local influence in measurement error models*, Biom. J. 38 (1996), pp. 829–841.
[28] S.-Y. Lee and N.-S. Tang, *Local influence analysis of nonlinear structural equation models*, Psychometrika. 69 (2004), pp. 573–592. DOI:10.1007/BF02289856
[29] S.-Y. Lee, B. Lu, and X.-Y. Song, *Assessing local influence for nonlinear structural equation models with ignorable missing data*, Comput. Stat. Data Anal. 50 (2006), pp. 1356–1377. http://www.sciencedirect.com/science/article/pii/S0167947304003883
[30] D. Mavridis and I. Moustaki, *The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data*, J. Comput. Graph. Stat. 18 (2009), pp. 1016–1034.
[31] T. Nakamura, *Corrected score function for errors-in-variables models: methodology and application to generalized linear models*, Biometrika. 77 (1990), pp. 127–137.
[32] A.S. Paulson, P. Roohan, and P. Sullo, *Some empirical distribution function tests for multivariate normality*, J. Stat. Comput. Simul. 28 (1987), pp. 15–30.
[33] W.-Y. Poon and Y.S. Poon, *Conformal normal curvature and assessment of local influence*, J. R. Stat. Soc. Ser. B Stat. Methodol. 61 (1999), pp. 51–61.

[34] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2016. https://www.R-project.org/
[35] A.R. Rasekh, *Local influence in measurement error models with ridge estimate*, Comput. Stat. Data Anal. 50 (2006), pp. 2822–2834. http://www.sciencedirect.com/science/article/pii/S0167947305001015
[36] A.R. Rasekh and N.R.J. Fieller, *Influence functions in functional measurement error models with replicated data*, Stat. J. Theor. Appl. Stat. 37 (2003), pp. 169–178.
[37] C.M. Russo, R. Aoki, and D. Leão-Pinto Jr, *Hypotheses testing on a multivariate null intercept errors-in-variables model*, Commun. Stat. Simul. Comput. 38 (2009), pp. 1447–1469. DOI:10.1080/03610910902972310
[38] S. Weisberg, *Applied Linear Regression*, Vol. 528, John Wiley and Sons, New Jersey, 2005.
[39] X. Zhang, H. Wang, Y. Ma, and R.J. Carroll, *Linear model selection when covariates contain errors*, J. Am. Stat. Assoc. 112 (2017), pp. 1553–1561.
[40] X.-P. Zhong, B.-C. Wei, and W.-K. Fung, *Influence analysis for linear measurement error models*, Ann. Inst. Stat. Math. 52 (2000), pp. 367–379.
[41] H.-T. Zhu and S.-Y. Lee, *Local influence for incomplete data models*, J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (2001), pp. 111–126.
[42] H. Zhu, J.G. Ibrahim, S. Lee, and H. Zhang, *Perturbation selection and influence measures in local influence analysis*, Ann. Stat. 35 (2007), pp. 2565–2588.