TECHNICAL

REPORT

0235

Kernel Weighted Influence Measures

N. Hens, M. Aerts, G. Molenberghs, H. Thijs, and G. Verbeke

IAP STATISTICS

NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Kernel Weighted Influence Measures

By NIEL HENS, MARC AERTS,
GEERT MOLENBERGHS, HERBERT THIJS
Center for Statistics, Limburgs Universitair Centrum,
Universitaire Campus, B-3590 Diepenbeek, Belgium
niel.hens@luc.ac.be, marc.aerts@luc.ac.be,
geert.molenberghs@luc.ac.be, herbert.thijs@luc.ac.be

and GEERT VERBEKE
Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35,
B-3000 Belgium
geert.verbeke@med.kuleuven.ac.be

Abstract

To asses the sensitivity of conclusions to model choices in the context of selection models for non-random dropout, several methods have been developed. None of them are without limitations. In this paper, a new method called kernel weighted influence is proposed. While global and local influence approaches look upon the influence of cases, this new method looks at the influence of types of observations. The basic idea is to combine the existing influence approaches with a nonparametric weighting scheme. The kernel weighted global influence offers a possible solution to the problem of masking, while the kernel weighted local influence can be seen as a tool to better understand the source of influence.

KEYWORDS: Local Influence, Global Influence, Kernel Weights, Missing Data, Sensitivity Analysis.

1 Introduction

In a longitudinal study, each unit is measured on several occasions. It is not unusual for some sequences of measurements to terminate early for reasons outside the control of the investigator, any unit so affected is often called a dropout. Little and Rubin (1987) make important distinctions between different missing values processes. A dropout process is said to be completely random (MCAR) if the dropout is independent of both unobserved and observed data and random (MAR) if, conditional on the observed data, the dropout is independent of the unobserved measurements; otherwise the dropout process is termed non-random (MNAR) or non-ignorable.

To represent such a model, Diggle and Kenward (1994) proposed a selection model which combines the measurement part with the missingness process. This model and other models trying to represent a non-random dropout mechanism, rely on strong and untestable assumptions. Not only the assumed distributional form can be misspecified but also the presence of influential observations can be of great importance. A well known method to investigate the influence of individual cases is case deletion (Cook and Weisberg 1982). This results in the global influence approach. A quite different approach is to perturb the model a bit and study the stability of the model, as is done by Lesaffre and Verbeke (1998) as an application of the local influence approach introduced by Cook (1986). In Thijs et al (2000), Molenberghs et al (2001) and Verbeke et al (2001), this method was used to investigate the influence of non-random missingness as part of a sensitivity analysis in the selection modelling framework. A thorough discussion can also be found in Verbeke and Molenberghs (2000).

One of the datasets discussed in the literature is the mastitis dataset. These data were initially used by Kenward (1998) for an informal sensitivity analysis. They were analyzed extensively with the local influence approach by Molenberghs et al (2001).

The influence analyses on the mastitis and other datasets, make it clear that the allocation of the possibly different sources of influence is still a burden. The related question on when to call a case influential (i.e., well defined cut off values) is still an open problem. In view of obtaining new insight in this matter, we introduce kernel weighted influence measures. We will illustrate the techniques on the mastitis dataset throughout this paper.

Our proposal is an extension of the two approaches of global and local influence. Instead of looking at cases, we are interested in looking at the influence of types of observations. To know why an observation is influential, one has to consider the characteristics of that observation. So, instead of wondering why this particular observation is influential, the question becomes which characteristics of this observation makes this type of observation influential. Therefore we will look at observations in the neighborhood of a case.

In the next section the mastitis dataset is introduced and described. The selection model of Diggle and Kenward and the global and local influence will briefly be reviewed in Section 3. The development and motivation of the kernel weighted influence measures is given in Section 4. This approach will be extended to a grid analysis in Section 5. In Section 6 a small simulation study is carried out.

2 The Mastitis Dataset

In this dataset the occurrence of the infectious disease of the udder, called mastitis, in dairy cows was studied. The milk yields in thousands of liters of 107 cows from a single herd in two consecutive years were available. In the first year all cows were supposedly free of mastitis and in the second year 27 cows became infected. Mastitis typically leads to a reduction in milk yield. There is a view among dairy scientists, widely held, that mastitis is more likely to occur in high yielding cows. It is however difficult to examine such a relationship due to the effects of mastitis.

Figure 1 shows a profile plot of the mastitis data.

FIGURE 1 ABOUT HERE

Looking at the different profiles in this figure, cows #4, #5 and #66 have a large increase in milk yield compared with the other cows. Cow #89 appears to have the largest decrement. Next to cow #66, cows #54, #69 and #53 are high yielding cows in both consecutive years.

Because some cows have a large reduction in milk yield and others exhibit a substantial increase, it is useful to look at the increments, i.e., the difference between the milk yield in the second year and the first year. In Figure 2, a scatterplot of the original data is given together with a plot of the increments against the first measurement.

FIGURE 2 ABOUT HERE

If we take a closer look at these two scatterplots, we can see that the cows mentioned above are located at the border of the data region. Whether or not these cows have a large influence on a statistical analysis is not clear without further investigation. This is the purpose of a sensitivity analysis. Kenward (1998) introduced a statistical model to analyze the mastitis data, a model that fits in the selection modelling framework, as introduced in the next section.

3 Influence Measures

This section summarizes parametric approaches to sensitivity analysis within the framework of selection models.

3.1 A Selection Model for Non-Random Dropout

Let us assume that for subject $i = 1, \cdots, N$, a sequence of responses Y_{ij} is measured at two occasions $j = 1, 2$. Let R_i be a missingness indicator and assume that y_{i1} is always observed. Then, $r_i = 1$ if y_{i2} is missing and $r_i = 0$ if y_{i2} is observed. The measurement part of the model of Diggle and Kenward (1994), which is in fact a linear mixed model, is characterized by

$$\mathbf{Y_i} = (Y_{i1}, Y_{i2}) \sim N(X_i\beta, \varSigma_i), \qquad i = 1, \ldots, N, \tag{1}$$

where β is a vector of fixed effects, X_i contains covariate values and $\varSigma_i$ is the covariance matrix. The missingness process is described by

$$\text{logit}[Pr(R_i = 1|y_{i1}, y_{i2})] = \psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}, \tag{2}$$

where $Pr(R_i = 1|y_{i1}, y_{i2})$ is the probability for the i^{th} subject to drop out, under the posited model. If ψ_2 differs from zero, the missingness process is non-random.

The measurement part used on the mastitis data is given by

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu \\ \mu + \Delta \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right], \tag{3}$$

where the covariance matrix expresses serial correlation.

The fit of this model on the mastitis data based on the assumption that the dropout process is MAR on the one hand and MNAR on the other hand (Diggle and Kenward 1994) is summarized in Table 1.

TABLE 1 ABOUT HERE

Testing $H_0 : \psi_2 = 0$ by means of a likelihood ratio test gives the value $G^2 = 5.11$, indicating some evidence against the MAR assumption. The high value of the test statistic does not at all mean that there are observations in the dataset which are missing not at random. It is also possible that this high value is due to misspecification of the distribution or even just the missingness process. An important question is then, whether some particular subjects are responsible for this behavior. Cook and Weisberg (1982) introduced a case deletion approach to investigate the influence of subjects. From their approach, several other methods were developed. The next two sections discuss global and local influence measures as applied on the mastitis data.

3.2 Global Influence

Let us introduce a weighted loglikelihood

$$l(\gamma; \mathbf{w}) = \sum_{j=1}^{N} w_j l_j(\gamma), \tag{4}$$

where $\mathbf{w} = (w_1, \ldots, w_N)$ is a vector of subject specific weights and $\ell_j(\gamma)$ represents the loglikelihood contribution of the j-th subject with γ the parameter vector containing all unknown parameters (from measurement and dropout model). The global influence measure CD_i compares the original loglikelihood $l(\hat{\gamma}; \mathbf{1})$ with the loglikelihood $l(\hat{\gamma}_{(-i)}; \mathbf{1})$, with $\hat{\gamma}$ and $\hat{\gamma}_{(-i)}$ the maximum likelihood estimators based on $l(\gamma; \mathbf{1})$ and $l(\gamma; \mathbf{w}_{(-i)})$ respectively, where $\mathbf{w}_{(-i)} = (1, \ldots, 1, 0, 1, \ldots, 1)$. The 0 is located at the i-th entry. Thus CD_i is given by

$$CD_i = 2(l(\hat{\gamma}; \mathbf{1}) - l(\hat{\gamma}_{(-i)}; \mathbf{1})). \tag{5}$$

A global influence analysis on the mastitis data, leads to influential cows #4, #5, #66 and #89, as shown in Figure 3.

FIGURE 3 ABOUT HERE

This is not surprising since cows #4, #5 and #66 have the largest increases in milk yield from year 1 to year 2 and cow #89 has the largest decrease in milk yield. Their behavior is thus different from the other cows. A full discussion is given by Molenberghs et al (2001).

There are two main disadvantages of the global influence method. The calculation of the Cook's distances requires $N+1$ model fits and the influence that can be ascribed to a specific cause is hard to assess, since by deleting a subject all sources of influence are lumped together, with little hope to disentangle them.

To overcome these limitations, *local* influence methods have been suggested.

3.3 Local Influence

The principle is to investigate how the results of an analysis are changed under infinitesimal perturbations of the model. Based on knowledge about mastitis, the increments appear to be important. A thorough motivation is given in Molenberghs et al (2001). Therefore a missingness process of the following form is considered.

$$\text{logit}[P(R_i = 1|Y_{i1}, Y_{i2})] = \psi_0 + \psi_1(Y_{i1} + Y_{i2}) + \omega_i(Y_{i2} - Y_{i1}), \tag{6}$$

where ω_i is a subject-specific weight, allowing the investigator to determine the local influence of one subject on the dropout model.

Cook (1986) proposed to measure the distance between $\widehat{\gamma}_\omega$, the maximum likelihood estimator of $\ell(\gamma|\omega)$ and $\widehat{\gamma}$, the maximum likelihood estimator of $\ell(\gamma)$, by the so-called likelihood displacement, defined by

$$LD(\omega) = 2(l(\hat{\gamma}|\omega_{\mathbf{0}}) - l(\hat{\gamma}_\omega|\omega_{\mathbf{0}})) \tag{7}$$

with $l(\gamma|\omega) = \sum_{i=1}^{N} l_i(\gamma|\omega_i)$ where $\ell_i(\gamma|\omega_i)$ denotes the i-the loglikelihood contribution associated with (7) and $\omega_{\mathbf{0}} = (0, \ldots, 0)$ the vector which corresponds to an MAR process. This approach takes into account the variability of $\widehat{\gamma}$. The geometric surface formed by the values of the graph $\xi(\omega) = (\omega, LD(\omega))$ gives the essential information about the influence of the perturbation scheme. Because of graphical limitations in dimensions higher than 2, Cook (1986) proposed to look at the normal curvatures $C_{\mathbf{h}}$ of $\xi(\omega)$ in $\omega_{\mathbf{0}}$, in the direction of some N-dimensional vector $\mathbf{h}$ of unit length.

Cook (1986) has shown that $C_{\mathbf{h}}$ can easily be calculated by

$$C_{\mathbf{h}} \;=\; 2 \left| \, \mathbf{h}' \, \mathit{\Delta}' \, \ddot{L}^{-1} \, \mathit{\Delta} \, \mathbf{h} \, \right|, \tag{8}$$

where $\mathit{\Delta}$ is a $(s \times N)$ matrix with $\mathbf{\Delta_i}$ as its i^{th} column, $\mathbf{\Delta_i}$ being the s dimensional vector defined by

$$\mathbf{\Delta_i} = \left. \frac{\partial^2 \ell_i(\gamma|\omega_i)}{\partial\omega_i \partial\gamma} \right|_{\gamma=\widehat{\gamma}, \omega_i=0} . \tag{9}$$

Further, $\ddot{L}$ denotes the $(s \times s)$ matrix of second order derivatives of $\ell(\gamma|\omega_0)$ with respect to γ, also evaluated at $\gamma = \widehat{\gamma}$.

One evident choice is the vector $\mathbf{h_i}$ containing 1 in the i^{th} position and 0 elsewhere, corresponding to a perturbation from the MAR model for the i^{th} subject in (7) only. This reflects the influence of allowing the i^{th} subject to drop out non-randomly, while the others can only drop out at random.

Calculating the local influences of the cows in the mastitis data, cows #4, #5 and #66 appear to be influential (see Figure 4). This is in agreement with the global influence analysis. Because the local influence looks at perturbations of the MNAR-parameter, while the global influence is based on case deletion, this was not to be expected a priori (Molenberghs et al, 2001). Kenward (1998) observed that cows #4 and #5, which show up in both analyses, are substantially different from the other cows by their large increment.

FIGURE 4 ABOUT HERE

If the dropout probabilities are considered, then cow #66 seems to have a large dropout probability compared with the other cows. Therefore, a perturbation of the MNAR-parameter will reflect this.

From both the global and local influence analyses it is clear that the location of the data is of great interest. Therefore a method to analyze sensitivity of types of observations might lead to a better comprehension of the influence measures and sensitivity analyses.

4 Kernel Weighted Influence Measures

The basic idea is to study the influence of types of observations, which are defined by neighborhoods centered at the observations (y_{1i}, y_{2i}, r_i). Here techniques from nonparametric smoothing methods can be used. Inspired by the well-known kernel estimators and density and regression estimators, kernel weights, as defined in the next section, can be used in the weighted loglikelihood (4) and the normal curvature (8) in order to derive new influence measures.

4.1 Kernel Weights

Influence measures such as the global influence and local influence approach are essentially based on cases. Our proposal is to extend these two approaches by looking in the neighborhood of the outcomes (y_{1i}, y_{2i}, r_i). Therefore, we introduce the following weights. If $r_i = 0$,

$$w_i(y_{1j}, y_{2j}, r_j) = \begin{cases} \frac{K^2(0)}{\mathrm{nd}_0} & r_j = 1 \\ \frac{K^2(0) - K(\frac{y_{1j}-y_{1i}}{g_1})K(\frac{y_{2j}-y_{2i}}{g_2})}{\mathrm{nd}_0} & r_j = 0 \end{cases} \tag{10}$$

and if $r_i = 1$,

$$w_i(y_{1j}, y_{2j}, r_j) = \begin{cases} \frac{K^2(0) - K(\frac{y_{1j}-y_{1i}}{g_1})K(0)}{\mathrm{nd}_1} & r_j = 1 \\ \frac{K^2(0)}{\mathrm{nd}_1} & r_j = 0 \end{cases} \tag{11}$$

where K is a gaussian kernel function, g_1 and g_2 are two possibly different bandwidths and r_j is the missingness indicator for subject j. The denominators nd_1 (if $r_i = 0$) and nd_2 (if $r_i = 0$) are equal to $\sum_{j=1}^{N} w_i(y_{1j}, y_{2j}, r_j)$. In this way the weights are standardized in the sense that they sum up to one. The motivation of the weights is as follows. If $r_i = 0$, (y_{1i}, y_{2i}) is a completer and all completers in the neighborhood get low weight. All other subjects get high weight, including the dropouts. If $r_i = 1$, y_{2i} is not observed and all dropouts in the neighborhood are given low weight, while all other subjects, including the completers, get high weight. This is graphically shown in Figure 5.

FIGURE 5 ABOUT HERE

4.2 Kernel Weighted Global Influence

To explore the neighborhood of the outcome (y_{1i}, y_{2i}, r_i), one can look at a vector $\mathbf{w_i}$ where the j^{th} component obtains weight $w_{ij} = w_i(y_{1j}, y_{2j}, r_j)$ as introduced in Section 4.1. This extension of the well known global influence approach is able to allocate groups of influential cases with similar outcomes, thus avoiding the problem of masking.

The choice of the bandwidth is one of the crucial points in this analysis. If a neighborhood contains a lot of observations, a large bandwidth would imply that all the observations in that neighborhood would be downweighted and the kernel weighted global influence measure would be large. If an observation is left out in the middle of a dense neighborhood, one expects that this would not have a large influence on the likelihood. Therefore, the bandwidth needs to be adjusted to the density of the point under consideration.

Consider (y_{1i}, y_{2i}, r_i), the datapoint of interest. If $r_i = 0$ the bandwidth is taken to be

$$h(y_{1i}, y_{2i}, r_i) = \frac{CK^2(0)}{\sum_{j,r_j=0} K(\frac{y_{1j}-y_{1i}}{g_1})K(\frac{y_{2j}-y_{2i}}{g_2})} \tag{12}$$

If $r_i = 1$ the bandwidth is taken to be

$$h(y_{1i}, y_{2i}, r_i) = \frac{CK^2(0)}{\sum_{j,r_j=1} K(\frac{y_{1j}-y_{1i}}{g_1})K(0)} \tag{13}$$

where C is a constant and g_1 and g_2 are two initially chosen bandwidths. Next to this adjustment, the normalizing denominator nd_1 and nd_2 assume that the total sample size remains unchanged.

A kernel weighted global influence analysis with initial bandwidths $g_1 = g_2 = 0.2$ and $g_1 = g_2 = 1.5$ on the mastitis data leads to Figures 6 and 7 respectively.

FIGURE 6 ABOUT HERE

FIGURE 7 ABOUT HERE

For both bandwidths the types of cows corresponding to #4, #5, #54, #66, #69 and #89 seem to have a large influence. From Figure 2 it is clear that these cows are those lying at the border of the region. Cows #54 and #69 were not found with the global influence. The profiles of these two cows are practically the same (Figure 1). The global influence did not identify these cows as influential due to masking. The maximum likelihood estimators $\hat{\gamma}_{(-\mathbf{54})}$, $\hat{\gamma}_{(-\mathbf{69})}$ as defined in Section 3.2 do not differ very much from $\hat{\gamma}$. In the kernel weighted global influence both cows get low weight and therefore, the shift in likelihood is detected. It is thus the type of observation which is important here. If we have a closer look to Figure 7, a second group of types of observations seems to be influential. This group corresponds to types of observations #7, #47 and #58, which are incomplete observations. These incomplete observations have the three highest y_1-values among the incompleters (Figure 1) and thus can be seen as outlying observations.

4.3 Kernel Weighted Local Influence

The local influence approach can be extended by looking at the direction $\mathbf{h_i}$ where the j^{th} component equals $1 - w_i(y_{1j}, y_{2j}, r_j)$ with w_i as defined in Section 4.1. This choice for $\mathbf{h_i}$ reflects the influence of allowing subjects in the neighborhood of the i-th subject to drop out non-randomly, while others, not within this neighborhood, can only dropout at random. This method provides new insights in the local influence of types of observations.

The choice of the bandwidths is crucial. Because the vector $\mathbf{h_i}$ is normalized, there is no need to have a density-adaptive bandwidth as in Section 4.2.

In the weighted local influence approach, applied on the mastitis data, one is interested in whether the probability of occurrence of mastitis is related to the yield that would have been observed had mastitis not occurred for a cow with certain characteristics. In Figure 8, a kernel weighted influence analysis for 6 different bandwidths is shown for the local influence analysis.

FIGURE 8 ABOUT HERE

For a larger bandwidth the left upper panel in Figure 8 suggests two groups of observations. The group with the highest influence is the group of completers, while the other group is the group of incompleters. If the bandwidth decreases, #66 shows up, as is shown in the right upper panel in Figure 8. For further decreasing bandwidths, #66 remains influential, while two other observations, #4 and #5, show up. The fact that #66 is dominantly present at several choices for the bandwidth, stresses the high degree of influence for this type of observations. The profile of #66 (Figure 1) is special in the sense that the milk yield in year 1 and year 2 are very high and so is the increase in milk yield. Types of observations with such a profile have a high dropout probability (Table 1), if they do not dropout they seem to be influential.

The kernel weighted influence approach has the additional advantage to allow for a grid-based influence analysis as explained in the next section.

5 Grid-Based Influence Measures

Instead of considering weights, centered at the datapoints (y_{1i}, y_{2i}, r_i), $i = 1, \ldots, 107$ of the mastitis dataset, we now consider weights centered at all points (y_1, y_2, r) on a one-$(r = 1)$ or two-dimensional grid $(r = 0)$ enclosing the full observed data range. These weights, given by the same expressions as in Section 4.1 but with the subscript i omitted everywhere, are used to calculate the kernel weighted global and kernel weighted local influence. The effect here is that we look at the dataset from the viewpoint of a gridpoint, which represents a possible type of observations which could have been in the sample. Graphical plots of the influence values as a function of y_1 (incompleters) or y_1 and y_2 (completers) can be used as exploratory sensitivity tools.

5.1 Grid-Based Kernel Weighted Global Influence

The two plots in Figure 9 show kernel weighted global influence values over a (y_1, y_2)-grid $[1, 9] \times [2, 12]$ in steps of 0.2. Again, as in Section 4.2, we used a density-adaptive bandwidth. The initial bandwidths g_1 and g_2 in (13) were chosen equal to 0.2 and 1.5 respectively.

FIGURE 9 ABOUT HERE

These plots show that, using the available information in the mastitis sample, certain types of observations are highly influential when modelled missing not at random in stead of missing at random. The peaks shown in Figure 9 confirm the results from Section 4.2. Indeed, a closer inspection of the first plot in Figure 9 reveals that the four highest peaks correspond to types of observations with characteristics similar to cows #4 and #5, to #54 and #69, to #66 and to #89.

The main structure of the second plot in Figure 9, based on a larger initial bandwidth, is essentially the same but the influence of observations at the border of the ellipsoidal area of datapoints gets more pronounced. Especially observations on that border, with Y_2 large, seem to be highly influential.

A similar grid analysis for the incompleters didn't show any highly influential cases.

The construction of such a grid-based global influence graph is very computer intensive due to the calculation of the numerous maximum likelihood estimates. This is not the case for a grid analysis based on kernel weighted local influence, which is computationally much simpler. This is illustrated in the next section.

Figure 9, thus, confirms what was seen in section 4.2.

5.2 Grid-Based Kernel Weighted Local Influence

Similar to the kernel weighted global influence, the kernel weighted local influence can be calculated over a grid. Using directions $\mathbf{h_i}$, similar to those in Section 4.3 but now based on weights centered at the grid points, a plot of the weighted local influence values can be constructed and might lead to additional insights. Since in this case the computations are rather simple and fast, we used a wider range and a finer grid. Moreover, it is feasible to consider a number of bandwidths. Figure 10 shows weighted local influence graphs for six different bandwidths.

FIGURE 10 ABOUT HERE

The main structure is essentially the same in each graph. If we have a closer look to the graphs for smaller bandwidths, the non-influential region is concentrated at the first principal component axis. The correlation between Y_1 and Y_2 is strongly positive, as can be seen in Figure 2. The types of observations which do not follow this main structure of the data, can be seen as potential outlying types of observations. Especially, types of observations

with low values for Y_1 and high values for Y_2 seem to be influential. The highest influence for each of the graphs in Figure 10 for decreasing bandwidth is reached for (y_1, y_2) equal to $(2.93, 9.34)$;$(2.93, 8.49)$;$(3.08, 7.72)$;$(3.62, 7.41)$; $(3.78, 7.18)$ and $(3.93, 7.10)$
respectively. A closer look to these highly influential types of observations and to the mastitis data shows that they are of the same type as observations #4 and #5. This confirms our findings in Section 4.3.

A plot (omitted from the text) of the grid-based kernel weighted local influence for different bandwidths for types of incomplete observations showed little influence compared with the types of complete observations. The influential types of incomplete observations, when present, are located in the center of the first measurement-range $(3.5, 7.5)$.

Figure 10 clarifies that the types of observations which are outlying are likely to be influential.

A simulation study for the kernel weighted influence measures can give us a better insight in the source of influence for both complete and incomplete types of observations. Computationally, it is not feasible to carry out a simulation study for the grid-based kernel weighted global influence. Therefore, we restrict ourselves to a simulation study for the grid-based kernel weighted local influence.

6 A Simulation Study

A small simulation study is carried out in order to obtain new insights in the different sources of influence. For this simulation study 100 similar datasets were generated. Each dataset consists of 107 subjects, each with two measurements generated from a bivariate normal distribution. Consider the following bivariate normal distribution, based on a compound symmetry covariance matrix:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 6.426 \\ 7.095 \end{pmatrix}, \begin{pmatrix} 2.865 & 2.324 \\ 2.324 & 2.865 \end{pmatrix}\right]. \tag{14}$$

The dropout process was generated according to the following model

$$\text{logit}[P(R_i = 1|Y_{i1}, Y_{i2})] = -3.379 + 0.387Y_{i1} + \psi_2 Y_{i2} \tag{15}$$

where ψ_2 is the MNAR-parameter. The choice for the parameters in both the measurement model and dropout process was based on a fit of this model with $\psi_2 = 0$ (MAR) on the mastitis data.

6.1 A First Setting

In a first simulation setting, 104 of the 107 subjects in each dataset were generated according to the process described above with ψ_2 equal to 0 (MAR).

Three subjects however were generated with $\psi_2 = -0.5$, so three observations were allowed to be missing not at random. To compare the additional influence of generating 3 subjects which are allowed to be missing not at random versus the situation where all subjects are allowed to be missing at random, an average influence measure was plotted in Figure 11 for the completers and in Figure 12 for the incompleters. This average influence measure is the difference between the average grid-based influence of 100 datasets with 3 subjects allowed to be missing not at random and the average grid-based influence of 100 datasets, where none of the subjects were allowed to be missing not at random.

FIGURE 11 ABOUT HERE

FIGURE 12 ABOUT HERE

If we consider the dropout structure in Figure 13 for both MAR ($\psi_2 = 0$) and MNAR ($\psi_2 = -0.5$) and relate this two the results shown in Figure 11, it becomes clear that completers which tend to have a large probability of dropping out, but do not, appear to be influential.

For the types of observations with a missing second measurement the largest influence is located at higher y_1 values as can be seen in Figure 12. Incomplete observations with a high dropout probability are influential.

FIGURE 13 ABOUT HERE

6.2 A Second Setting

In a second simulation setting, the presence of subjects missing not at random is invoked by taking 100 datasets generated under MAR ($\psi_2 = 0$) as above, but now all data, with a second measurement higher than 8.5, are set to be missing.

In Figure 14, the average influence measure of the completers of 100 datasets is shown. We will refer to these datasets generated under MAR as the reference datasets.

FIGURE 14 ABOUT HERE

The plot of the average influence of the completers of the reference datasets versus the grid has a particular shape. There is very low or no influence for data along the first principal component axis due to the high correlation ($\rho_{Y_1,Y_2} = 0.80$) between Y_1 and Y_2. When we move away from this axis the average influence increases. This indicates that outlying types of observations, not following the main pattern in the data, are influential. To see what the effect of invoking MNAR-dropout is on the completers, we leave out all observations in these datasets with a Y_2-measurement higher than 8.5 and calculate the average kernel weighted local influence again.

FIGURE 15 ABOUT HERE

The average influence of the completers under such a MNAR dropout process is shown in Figure 15, which indicates that dropout due to this MNAR mechanism has a large change in influence for types of completers with a high Y_1-measurement and a low Y_2-measurement. This is confirmed by the contour plots, shown in Figure 16.

FIGURE 16 ABOUT HERE

The larger influence of observations with a high Y_1-measurement and a low Y_2-measurement is not surprising. In Figure 17 a scatterplot of the completers is given.

FIGURE 17 ABOUT HERE

If we consider the structure of the data, we know that observations with a high value for Y_1 are more likely to be missing due to the underlying MAR-mechanism (Figure 13). Combined with the MNAR-mechanism we invoked in this setting, we especially obtain complete observations with a low Y_2-measurement. The correlation indicates that, among these types of observations, the ones with a low Y_1-measurement follow the correlation structure of the data. The ones with a high Y_1-measurement do not follow this structure and therefore they can be seen as outlying types of observations. Their influence is rather high compared with the other types of observations.

Looking at the incompleters in Figure 18 one can see that there is a large change in influence on the incompleters. The highest average influence for the incompleters of the reference datasets was reached for $Y_1 = 8.5$, considering the MNAR-mechanism there is a shift towards $Y_1 = 9.75$. Not only this shift can be seen, but also the overall average influence increases. This indicates that the presence of types of observations which are left out non-randomly seem to have a large influence.

FIGURE 18 ABOUT HERE

Other simulation settings (such as larger sample sizes) confirm these results, the main idea is illustrated here and therefore these other simulations are omitted from this paper.

7 Conclusion and Final Remarks

The presence of influential observations in a dataset can disturb model fitting and model building thoroughly. Therefore it is essential to perform a sensitivity analysis when doing a data analysis. In this paper we introduced some new exploratory and graphical tools for sensitivity analysis, combining parametric global and local influence measures with nonparametric smoothing weights. These methods provide new insights in the influence of certain types of observations and offer a nice solution to the problem of masking. The presentation here has been focusing on the setting of two (repeated) measurements but, using more dimensional kernels or other higher dimensional distance measures, the method can be extended to three or more measurements. This is a topic of future research.

Acknowledgments

We gratefully acknowledge support form the Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

Cook, R.D. (1986) Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133–169.

Cook, R.D. and Weisberg, S. (1982) Residuals and influence in regression. *New York: Chapman and Hall.*

Diggle, P.J. and Kenward, M.G. (1994) Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.

Kenward, M.G. (1998) Selection models for repeated measurements with nonrandom dropout: an illustration of sensitivity. *Statistics in Medicine*, **17**, 2723-2732.

Lesaffre, E. and Verbeke, G. (1998) Local influence in linear mixed models. *Biometrics*, **54**, 570-582.

Little, R.J.A. & Rubin, D.B. (1987) Statistical Analysis with Missing Data. *New York: Wiley.*

Molenberghs, G., Verbeke, G., Thijs, T., Lesaffre, E. and Kenward, M.G. (2001) Influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, **37**, 93–113.

Thijs, H., Molenberghs, G. and Verbeke, G. (2000) The Milk Protein Trial: Influence analysis of the Dropout Process. *Biometrical Journal*, **42**, 1–30.

Verbeke, G. and Molenberghs, G. (2000) Linear Mixed Models for Longitudinal Data. *New York: Springer Verlag.*

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. and Kenward, M.G. (2001) Sensitivity Analysis for Non-Random Dropout: A Local Influence Approach. *Biometrics*, **57**, 7–14.

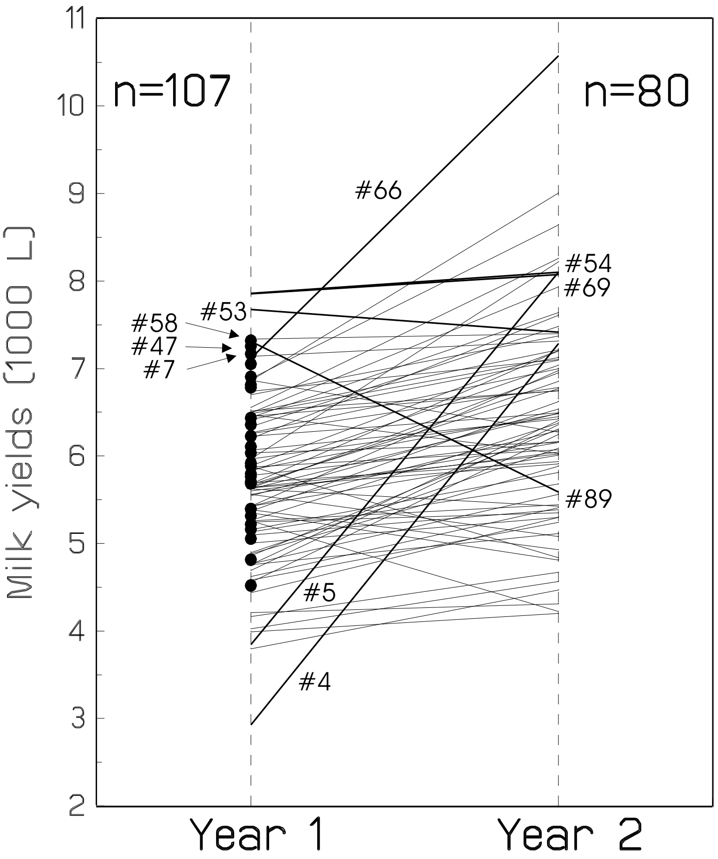


Figure 1: Profile plot of the mastitis dataset.

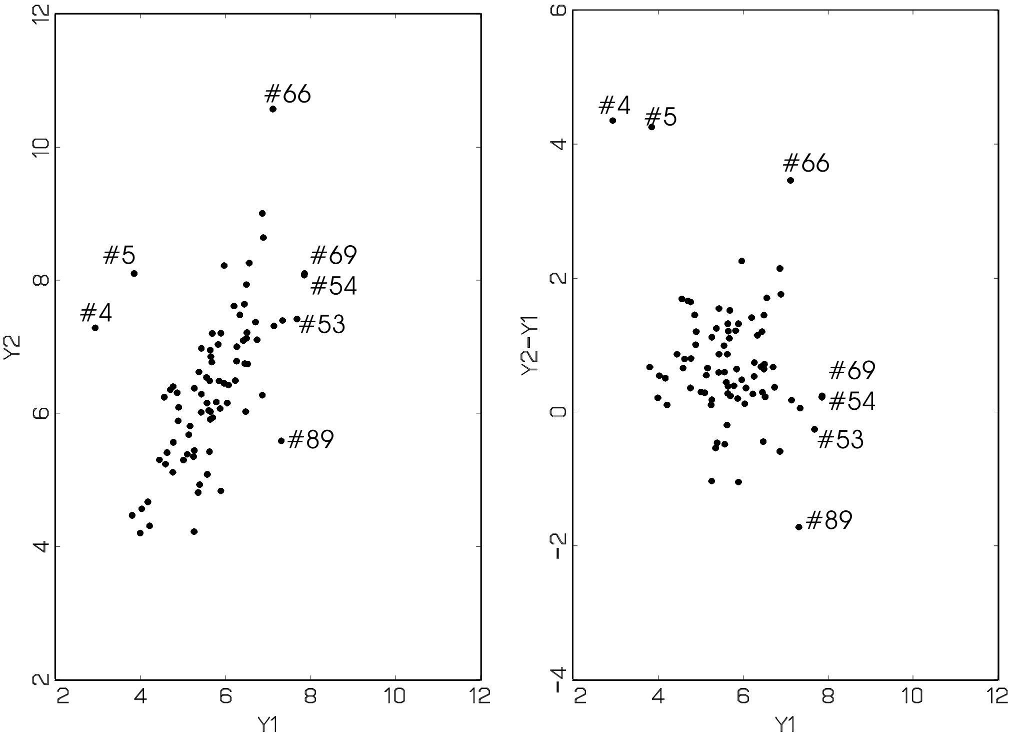


Figure 2: Scatter plot of the mastitis dataset. In the left panel the milk yield for year 2 was plotted versus the milk yield at year 1. In the right panel the increase in milk yield from year 1 to year 2 was plotted versus the milk yield at year 1.

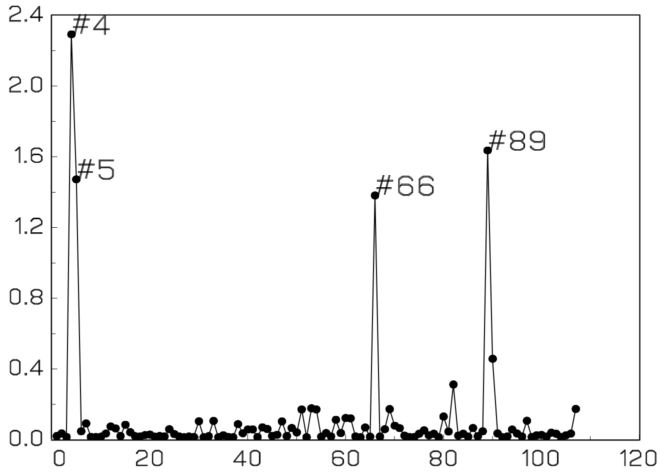


Figure 3: Influential subjects of the mastitis data based on the global influence measure.

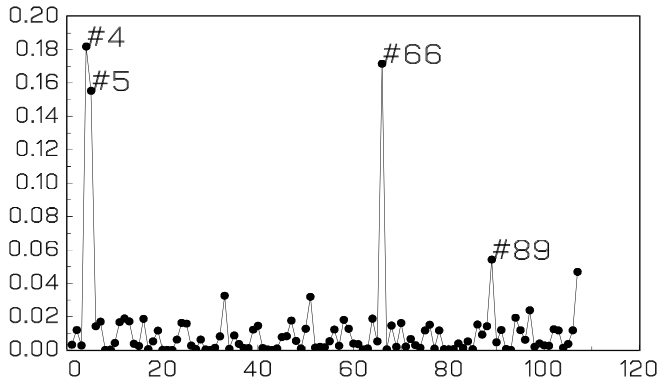
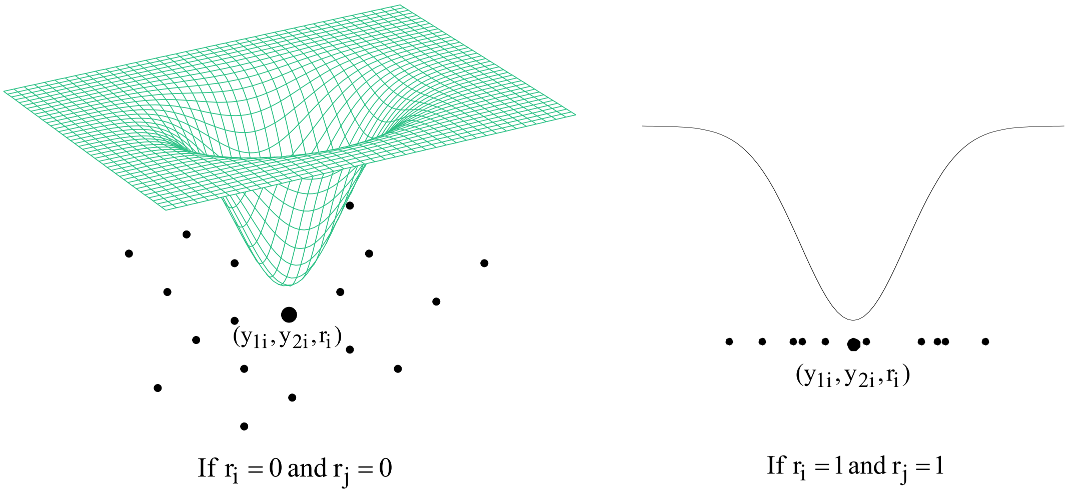


Figure 4: Influential subjects of the mastitis dataset using the local influence measure.



If $r_i = 0$ and $r_j = 0$

If $r_i = 1$ and $r_j = 1$

Figure 5: Shape of the weights. On the left hand side the weights are shown for the situation $r_i = 0$ and $r_j = 0$ (completers), while on the right hand side the weights are shown for the situation $r_i = r_j = 1$ (incompleters).

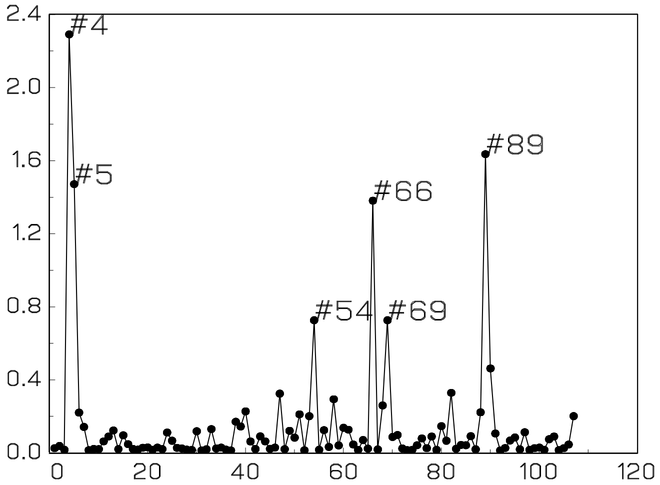


Figure 6: Influential subjects of the mastitis data for the kernel weighted global influence with initial bandwidths $g_1 = g_2 = 0.2$.

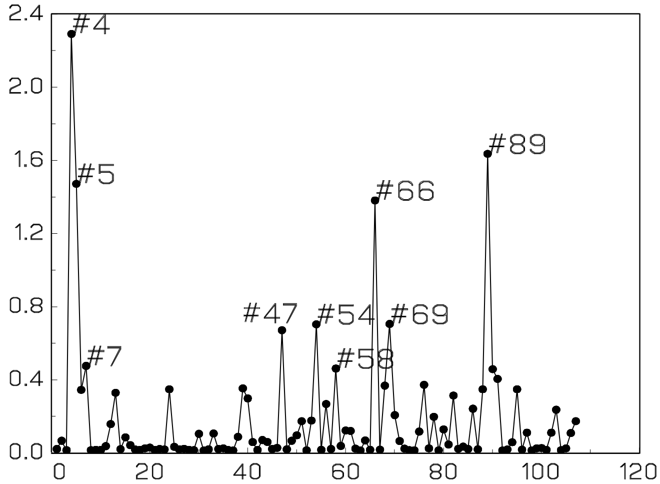


Figure 7: Influential subjects of the mastitis data for the kernel weighted global influence with initial bandwidths $g_1 = g_2 = 1.5$.

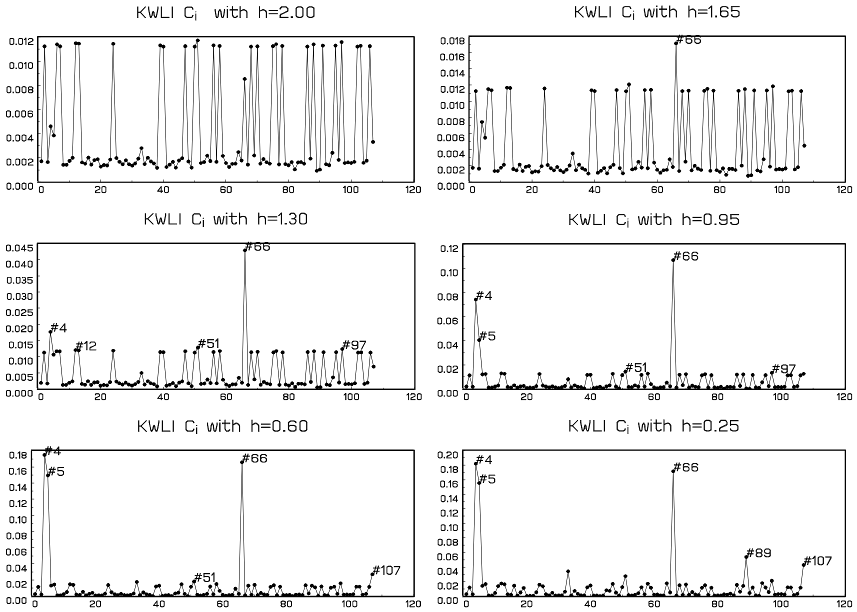


Figure 8: Influential subjects of the mastitis data for the kernel weighted local influence (increments) with different bandwidths $g_1 = g_2 = h$.

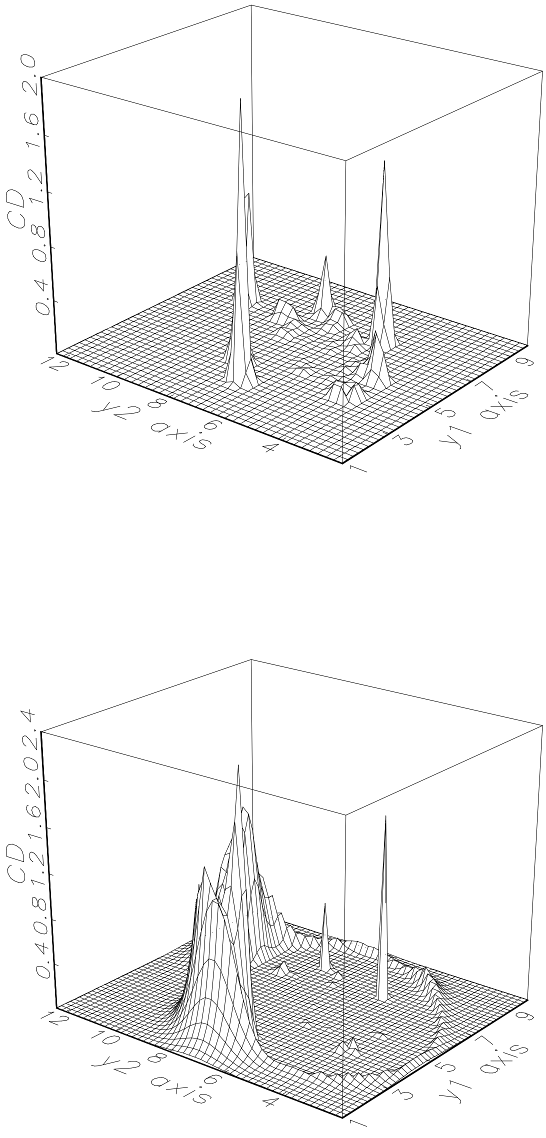


Figure 9: Kernel weighted global influence graph over a grid of completers with density-adaptive bandwidths initially equal to 0.2 (upper panel) and 1.5 (lower panel).

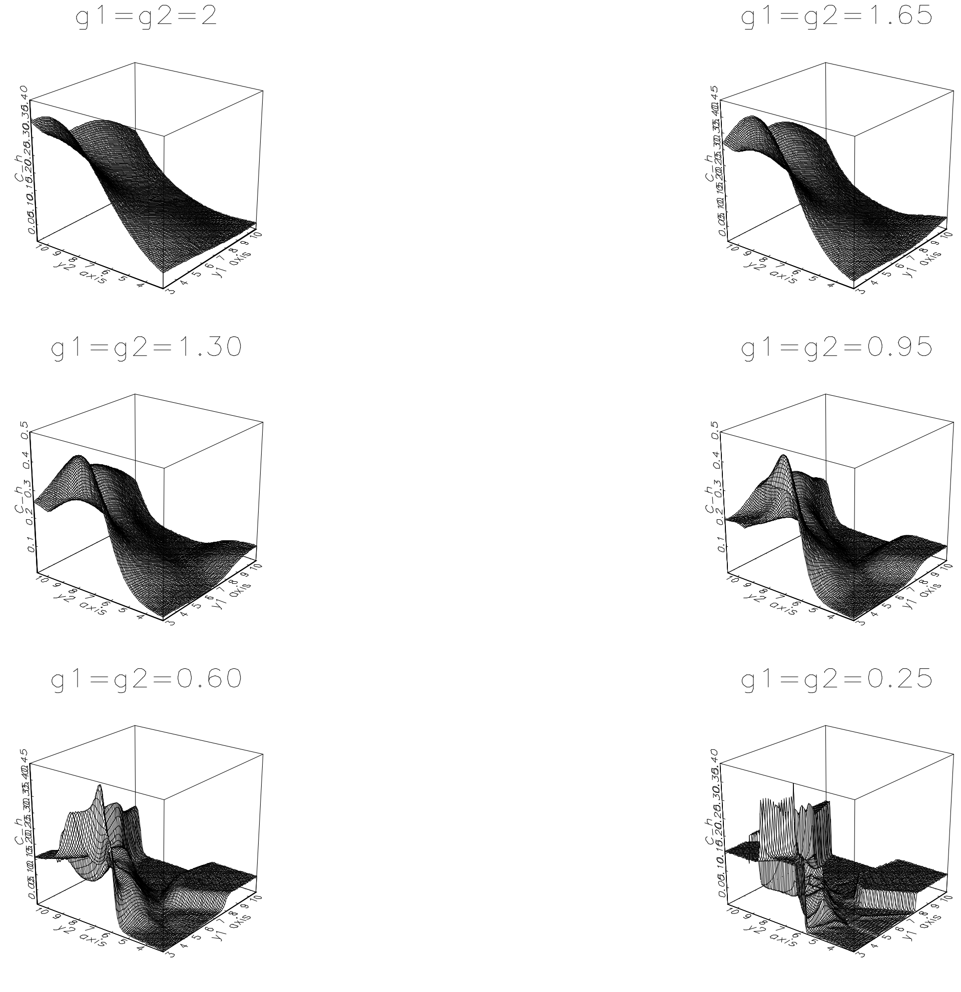


Figure 10: Kernel weighted local influence graphs over a grid of completers for several bandwidths $g_1 = g_2$.

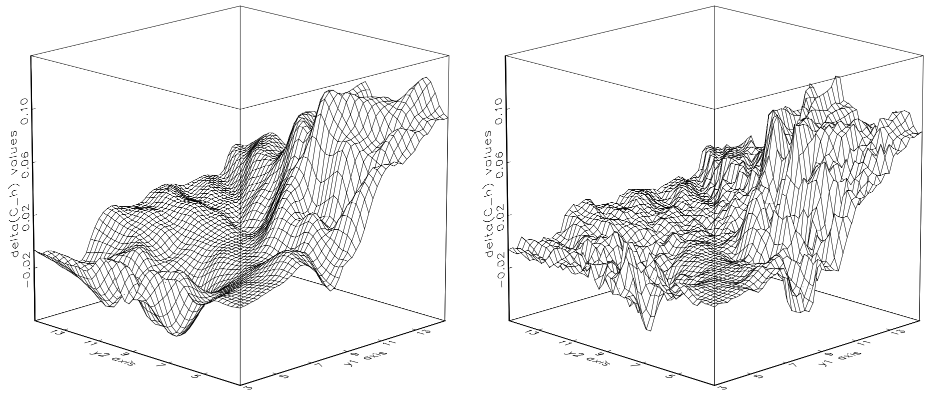


Figure 11: A figure of the relative average gain in influence of the completers when generating 3 subjects under MNAR. The bandwidths used are respectively equal to 1 and 0.5.

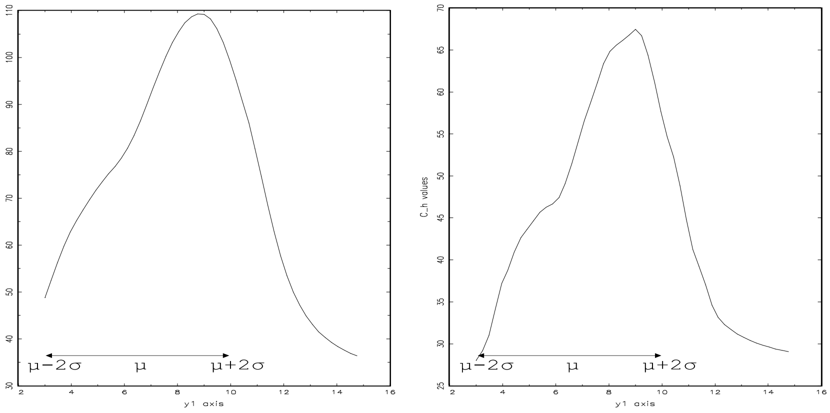


Figure 12: A figure of the relative average gain in influence of the incompleters when generating 3 subjects under MNAR. The bandwidths used are respectively equal to 1 and 0.5. μ and σ denote the mean and standard deviation of Y_1.

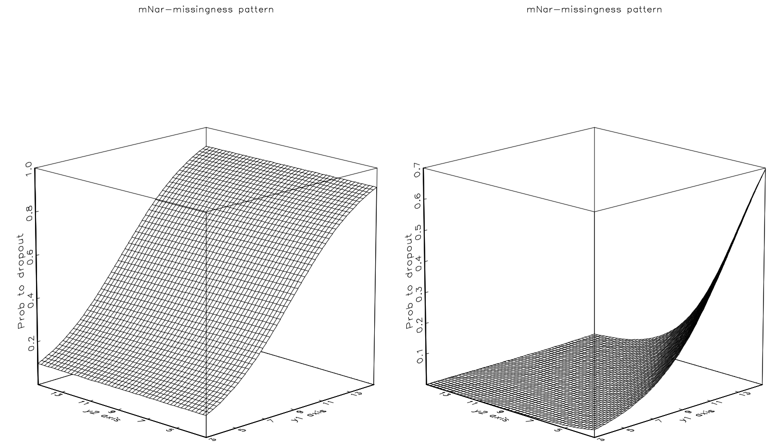


Figure 13: Plot of the probability of dropout. On the left hand side the dropout probability under MAR is shown, while on the right hand side the dropout probability under MNAR is shown.

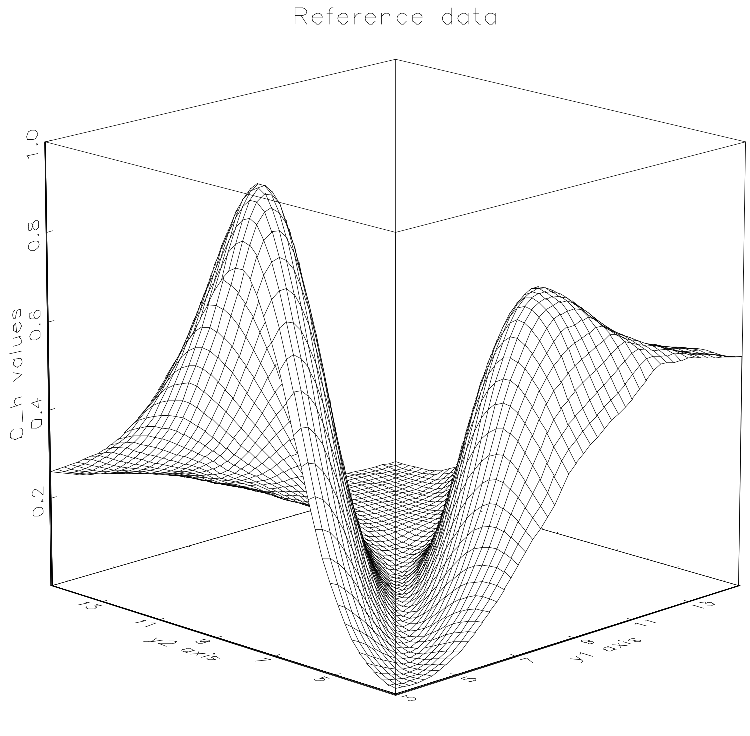


Figure 14: The average kernel weighted local influence for the completers of the 100 reference datasets

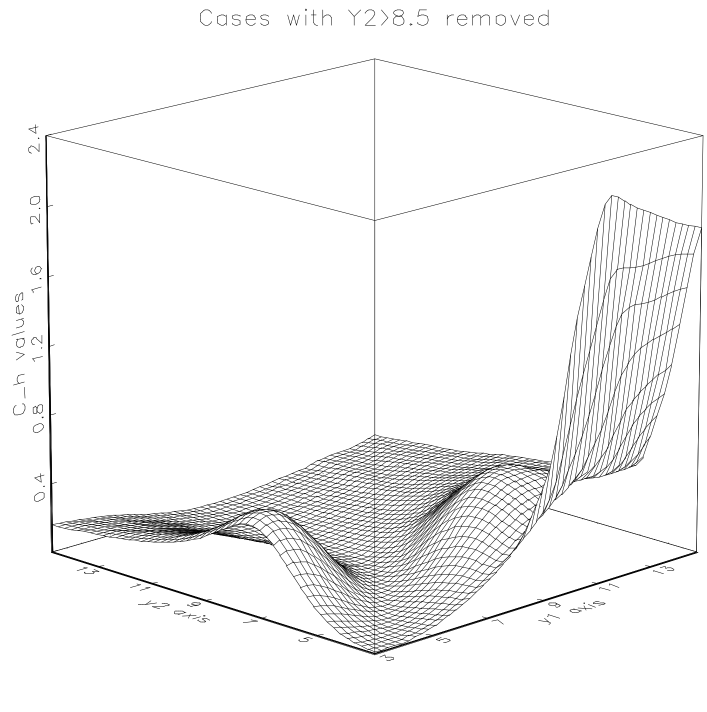


Figure 15: Kernel weighted local influence for the completers of the 100 complete datasets with MNAR dropouts for $Y_2 > 8.5$

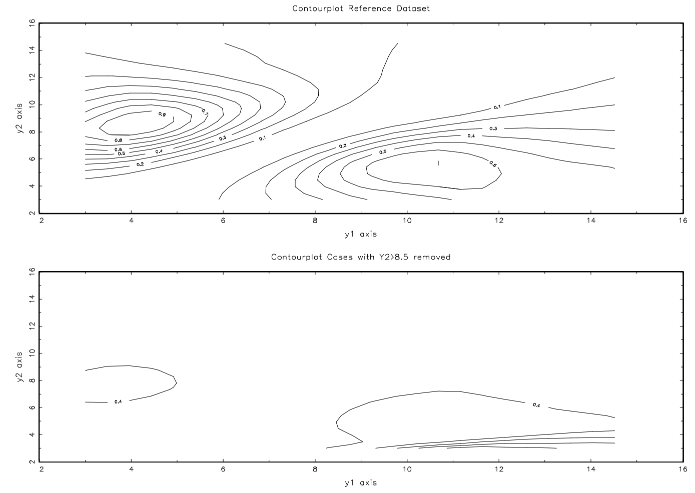


Figure 16: The contour plots of the kernel weighted local influence for the completers of 100 complete datasets and the completeres of the 100 datasets with MNAR dropouts for $Y_2 > 8.5$

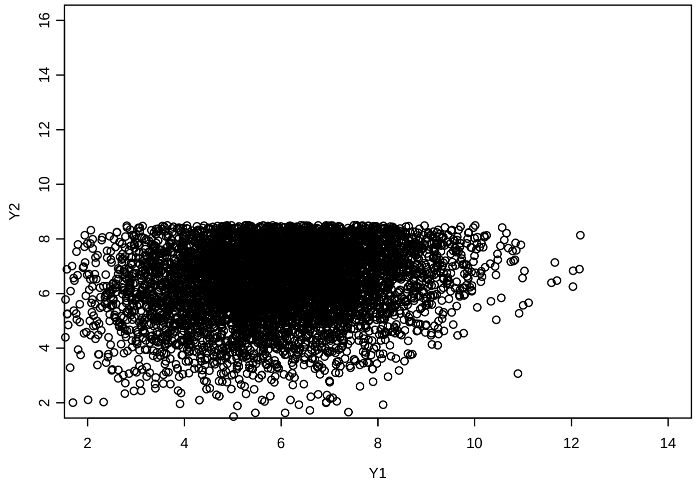


Figure 17: A scatterplot for all simulated datasets with MNAR dropouts for $Y_2 > 8.5$

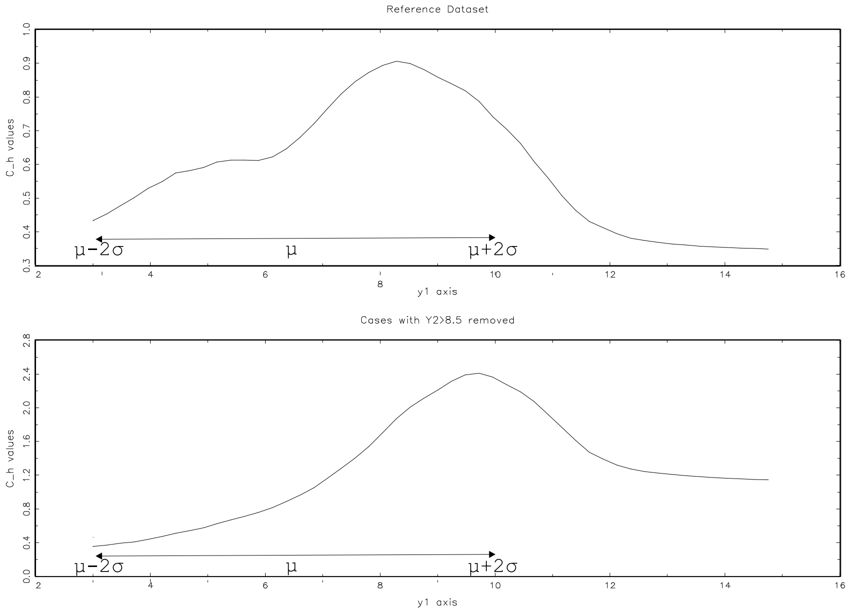


Figure 18: The figures of kernel weighted local influence for the incompleters of the complete dataset and the incompleters of the datasets with MNAR dropouts for $Y_2 > 8.5$

Effect	Parameter	Random Dropout	Non-Random Dropout
Measurement Model			
Intercept	μ	5.77(0.09)	5.77(0.09)
Time effect	Δ	0.72(0.11)	0.33(0.14)
First variance	σ_1^2	0.87(0.12)	0.87(0.12)
Second variance	σ_2^2	1.30(0.20)	1.61(0.29)
Correlation	ρ	0.58(0.07)	0.48(0.09)
Dropout Model			
Intercept	ψ_0	-2.65(1.45)	0.37(2.33)
First measurement	ψ_1	0.27(0.25)	2.25(0.77)
Second measurement	ψ_2	0	-2.54(0.83)
-2 loglikelihood		280.02	274.91

Table 1: Parameter estimates of the selection model fitted on the mastitis dataset.