

## Kernel weighted influence measures

Peer-reviewed author version

HENS, Niel; AERTS, Marc; MOLENBERGHS, Geert; THIJS, Herbert & VERBEKE, Geert (2005) Kernel weighted influence measures. In: COMPUTATIONAL STATISTICS & DATA ANALYSIS, 48(3). p. 467-487.

DOI: 10.1016/j.csda.2004.02.010

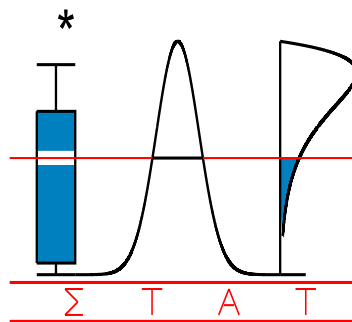
Handle: <http://hdl.handle.net/1942/2055>

# T E C H N I C A L R E P O R T

0465

## KERNEL WEIGHTED INFLUENCE MEASURES

HENS, N., AERTS, M., MOLENBERGHS, G., THIJS, H. and G. VERBEKE



I A P S T A T I S T I C S  
N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

# Kernel Weighted Influence Measures

Niel Hens <sup>\*,a</sup> Marc Aerts <sup>a</sup> Geert Molenberghs <sup>a</sup> Herbert Thijs <sup>a</sup>  
Geert Verbeke <sup>b</sup>

<sup>a</sup>*Center For Statistics, Limburgs Universitair Centrum, Universitaire Campus,  
Building D, B-3590 Diepenbeek, Belgium*

<sup>b</sup>*Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B-3000  
Belgium*

---

## Abstract

To assess 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. 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.

*Key words:* Local Influence, Global Influence, Kernel Weights, Missing Data, Sensitivity Analysis, Weighted Likelihood

---

## 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

---

\* Corresponding author. Tel. +32-11-26-8232; fax: +32-11-26-8299  
*Email address:* niel.hens@luc.ac.be (Niel Hens).

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. This new exploratory and graphical tool supplements many other tools for sensitivity analysis and can contribute in obtaining further insights in the mechanisms generating missing data.

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

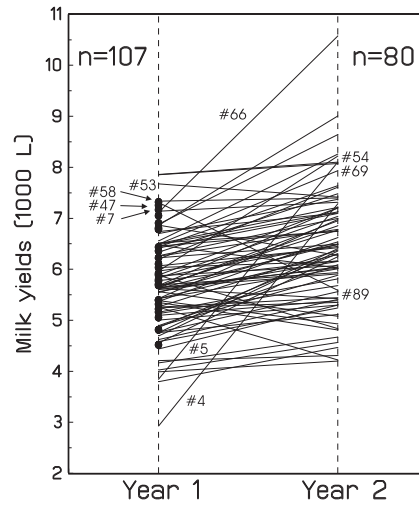


Fig. 1. Profile plot of the mastitis dataset.

carried out. In this paper we restrict attention to the case of two measurements of each subject. How the method can be extended to the general case of more than two time points is briefly sketched in Section 7.

## 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.

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.

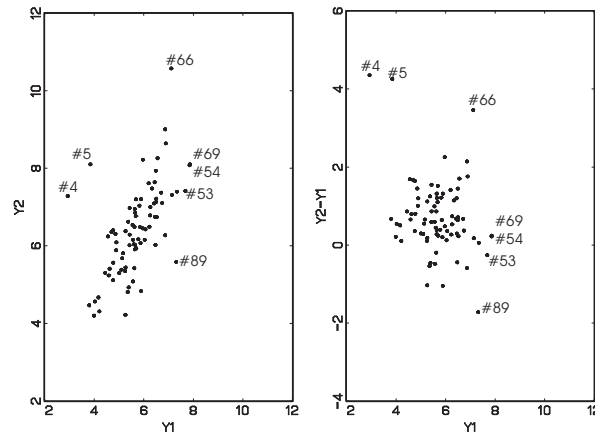


Fig. 2. Scatter plot of the mastitis dataset. In the left panel the milk yield for year 2 was plotted versus the milk yield at year 1. In the right panel the increase in milk yield from year 1 to year 2 was plotted versus the milk yield at year 1.

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. Are these specific cows having a large influence on a statistical analysis and are there any other cases with high influence? Getting more insights in such questions is the purpose of a sensitivity analysis. Special attention goes to cows #54 and #69, having almost identical measurements. It is known that, in classical regression models without missingness, most influence measures are not able to detect such cases, because they mask each other (see e.g. Ryan 1997). One of the main objectives is to study to which extent the influence measures introduced by Molenberghs et al (2001) suffer from the same deficiency; and to propose modified versions of these influence measures which deal with it. Another objective is to extend the methodology to measure the influence of ‘types’ of observations, not really included in the sample but represented by clusters of neighboring observations.

Kenward (1998) introduced a statistical model to analyze the mastitis data, a model that fits in the selection modelling framework. We briefly describe it 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  $i = 1, \dots, N$ , 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), applied to the mastitis data, is characterized by, for  $i = 1, \dots, N$ ,

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N} \left[ \begin{pmatrix} \mu \\ \mu + \Delta \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right]. \quad (1)$$

and the missingness process is described by

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

where  $Pr(R_i = 1|y_{i1}, y_{i2})$  is the probability for the  $i^{\text{th}}$  subject to drop out, under the posited model. If  $\psi_2$  differs from zero, the missingness process is non-random.

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.

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

$$l(\boldsymbol{\gamma}; \boldsymbol{w}) = \sum_{j=1}^N w_j l_j(\boldsymbol{\gamma}), \quad (3)$$

Table 1

Parameter estimates (and standard errors) of the selection model fitted on the mastitis dataset.

Effect	Parameter	Random Dropout	Non-Random Dropout
<u>Measurement Model</u>			
Intercept	$\mu$	5.77(0.09)	5.77(0.09)
Time effect	$\Delta$	0.72(0.11)	0.33(0.14)
First variance	$\sigma_1^2$	0.87(0.12)	0.87(0.12)
Second variance	$\sigma_2^2$	1.30(0.20)	1.61(0.29)
Correlation	$\rho$	0.58(0.07)	0.48(0.09)
<u>Dropout Model</u>			
Intercept	$\psi_0$	-2.65(1.45)	0.37(2.33)
First measurement	$\psi_1$	0.27(0.25)	2.25(0.77)
Second measurement	$\psi_2$	0	-2.54(0.83)
-2 loglikelihood		280.02	274.91

where  $\mathbf{w} = (w_1, \dots, w_N)$  is a vector of subject specific weights such that  $\sum_{i=1}^N w_i = N$  (reflecting an effective total sample of size  $N$ ) and  $l_j(\boldsymbol{\gamma})$  represents the loglikelihood contribution of the  $j$ -th subject with  $\boldsymbol{\gamma}$  the parameter vector containing all unknown parameters (from measurement and dropout model). Denote  $\hat{\boldsymbol{\gamma}}$  the maximum likelihood (ML) estimator of the unweighted likelihood, corresponding to the weight vector  $\mathbf{1} = (1, \dots, 1)$ , and  $\hat{\boldsymbol{\gamma}}_w$  the ML estimator based on the weighted likelihood (3).

Define

$$CD(\mathbf{w}) = 2\{l(\hat{\boldsymbol{\gamma}}; \mathbf{1}) - l(\hat{\boldsymbol{\gamma}}_w; \mathbf{1})\}, \quad (4)$$

as a measure for the distance between the ML estimator  $\hat{\boldsymbol{\gamma}}$  and the weighted ML estimator  $\hat{\boldsymbol{\gamma}}_w$ . The global influence measure (Molenberghs et al 2001)

$$CD_i = CD(\mathbf{w}_{(-i)}), \quad (5)$$

compares  $\hat{\boldsymbol{\gamma}}$  to  $\hat{\boldsymbol{\gamma}}_{(-i)}$ ; the latter is the weighted ML estimator using weight vector  $\mathbf{w}_{(-i)} = N/(N-1) \times (1, \dots, 1, 0, 1, \dots, 1)$  with the 0 at the  $i$ -th entry.

A global influence analysis on the mastitis data, leads to influential cows #4, #5, #66 and #89, as shown in Figure 3. This is not surprising since cows #4, #5 and #66 have the largest increases in milk yield from year 1 to year



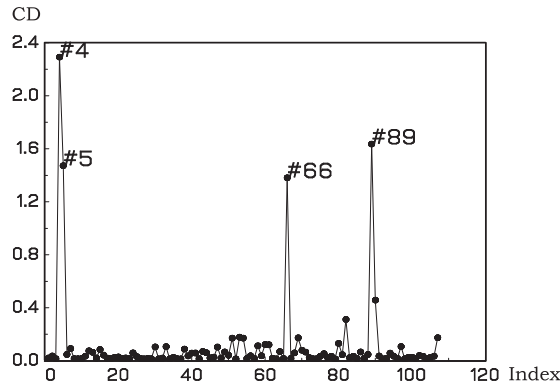


Fig. 3. Influential subjects of the mastitis data based on the global influence measure.

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). But apparently cows #54 and #69 are not suggested to be influential by the global influence measure  $CD_i$ .

A main disadvantage of global influence measures is that the influence that can be ascribed to a specific case is hard to assess, since by deleting a subject all sources of influence are lumped together, with little hope to disentangle them. This was the main motivation to look at local influence methods.

### 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.

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

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

Let  $l_i(\gamma|\phi_i)$  denote the  $i$ -th loglikelihood contribution of the  $i$ -th subject, associated with missingness process (6) and let  $l(\gamma|\phi) = \sum_{i=1}^N l_i(\gamma|\phi_i)$  denote the total loglikelihood with  $\phi = (\phi_1, \dots, \phi_N)$ . The vector  $\phi_0 = (0, \dots, 0)$  corresponds to an MAR process. Cook (1986) proposed to measure the distance between  $\hat{\gamma}_\phi$ , the ML estimator based on  $l(\gamma|\phi)$  and  $\hat{\gamma}_0$ , the ML estimator based on  $l(\gamma|\phi_0)$ , by the so-called likelihood displacement, defined by

$$LD(\phi) = 2\{l(\hat{\gamma}_0|\phi_0) - l(\hat{\gamma}_\phi|\phi_0)\}. \quad (7)$$

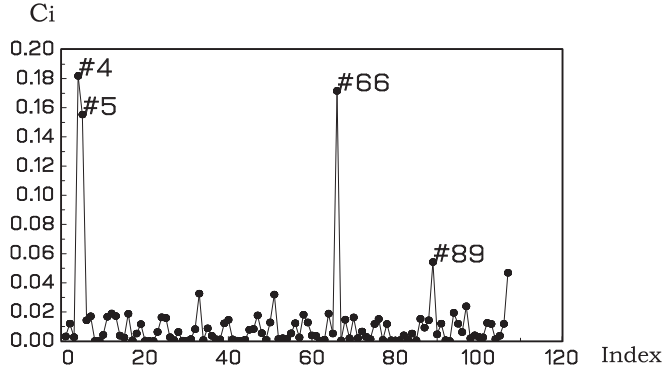


Fig. 4. Influential subjects of the mastitis dataset using the local influence measure.

This approach takes into account the variability of  $\hat{\gamma}$ . The geometric surface formed by the values of the graph  $\xi(\phi) = (\phi, LD(\phi))$  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  $\xi(\phi)$  at  $\phi_0$ , 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

$$C_{\mathbf{h}} = 2 \left| \mathbf{h}' \Delta' \ddot{\mathbf{L}}^{-1} \Delta \mathbf{h} \right|, \quad (8)$$

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

$$\Delta_i = \frac{\partial^2 l_i(\gamma|\phi_i)}{\partial \phi_i \partial \gamma} \bigg|_{\gamma=\hat{\gamma}_0, \phi_i=0}. \quad (9)$$

Further,  $\ddot{\mathbf{L}}$  denotes the  $(s \times s)$  matrix of second order derivatives of  $l(\gamma|\phi_0)$  with respect to  $\gamma$ , also evaluated at  $\gamma = \hat{\gamma}_0$ . One evident choice for  $\mathbf{h}$  is the vector  $\mathbf{h}_i$  containing 1 in the  $i^{\text{th}}$  position and 0 elsewhere, corresponding to a perturbation from the MAR model for the  $i^{\text{th}}$  subject in (7) only. The measure  $C_{\mathbf{h}_i}$  reflects the influence of allowing the  $i^{\text{th}}$  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.

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 analysis 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 for density and regression estimation (Wand and Jones 1995), we propose the use of a kernel based choice for the weight vector  $\mathbf{w}$  in the global measure (4) and for the direction vector  $\mathbf{h}$  in the local measure (8).

### 4.1 Kernel Weighted Global Influence

Influence measures such as the global influence and local influence approach are essentially based on the influence of single cases. The global measure  $CD_i$  quantifies the change in the parameter estimates when including or excluding the  $i$ -th case; the local measure  $C_{\mathbf{h}_i}$  reflects the influence of allowing the  $i$ -th subject to drop out non-randomly. We extend these two approaches by considering a neighborhood  $N(i)$  of  $(y_{1i}, y_{2i}, r_i)$  defined by kernel functions (see e.g. Wand and Jones 1995). Let  $K$  be a density function and  $g_1$  and  $g_2$  two so-called bandwidth parameters.

The neighborhood  $N(i)$  of observation  $i$  is characterized by the values of the product (or multiplicative) kernel

$$K\left(\frac{y_{1j} - y_{1i}}{g_1}\right)\{K\left(\frac{y_{2j} - y_{2i}}{g_2}\right)\}^{(1-r_i)}I(r_j = r_i), \quad (10)$$

for  $j = 1, \dots, N$ , where  $I(r_j = r_i)$  equals 1 if  $r_j = r_i$  and 0 otherwise. The first two factors in the definition of (10) are typical kernels for continuous variables and the indicator function can be considered as a kernel for a categorical variable. Taking the product of one-dimensional kernels is a typical simple way to characterize multivariate observations in the neighborhood of a certain

observation (see e.g. Wand and Jones 1995). The exponent of the second factor expresses the possible missingness of the second measurement  $y_2$ .

First consider the case observation  $i$  is complete ( $r_i = 0$ ). Complete observations ( $r_j = r_i = 0$ ) with values close to  $K^2(0)$  (the upper limit) are close neighbors of the  $i$ th observation; observations at a further distance have values for (10) close to 0 (the lower limit). Incomplete observations ( $r_j = 1$ ) get value 0. In case observation  $i$  is incomplete ( $r_i = 1$ ), the interpretation is essentially the same focusing on the first factor, now having a maximal value  $K(0)$  for the closest neighbors (identical observations).

This leads to the following definition: the kernel based weight vector  $\mathbf{w}_{(-N(i))}$  is the vector of length  $N$  with elements, for  $j = 1, \dots, N$ ,

$$(\mathbf{w}_{(-N(i))})_j = \left[ K(0)\{K(0)\}^{(1-r_i)} - K\left(\frac{y_{1j} - y_{1i}}{g_1}\right)\{K\left(\frac{y_{2j} - y_{2i}}{g_2}\right)\}^{(1-r_i)} I(r_j = r_i) \right] / D. \quad (11)$$

The denominator  $D$  is a normalization constant assuring that  $\sum_{i=1}^N (\mathbf{w}_{(-N(i))})_j = N$ .

Define the kernel weighted global influence measure of the  $i$ th observation ( $y_{1i}, y_{2i}, r_i$ ) as

$$CD_{N(i)} = CD(\mathbf{w}_{(-N(i))}). \quad (12)$$

It measures the discrepancy between the ML parameter estimator including or excluding the neighborhood  $N(i)$  as indicated by the weight vector  $\mathbf{w}_{(-N(i))}$ . The weights are shown graphically in Figure 5. For bandwidths  $g_1$  and  $g_2$  tending to 0 and in case all observations are different (no ties), the weight vector  $\mathbf{w}_{(-N(i))}$  converges to  $\mathbf{w}_{(-i)}$ . In case there are ties (or very close neighbors), the method contrasts the parameter estimates including or excluding these particular ties (or very close neighbors) for bandwidths tending to 0 (or very small). So the kernel weighted influence measure (12) is able to allocate groups of influential cases with similar outcomes, thus avoiding the problem of *masking*. Masking refers to the existence of a close cluster of influential data points such that deleting a single point will cause little effect (see e.g. Ryan 1997).

As the method is intended as an exploratory and graphical tool, the influence of neighborhoods  $N(i)$ , characterizing a certain type of observation, is explored by considering a series of bandwidth values. But, from our experience, the bandwidth needs to be adjusted to the data density at the observation  $i$  under

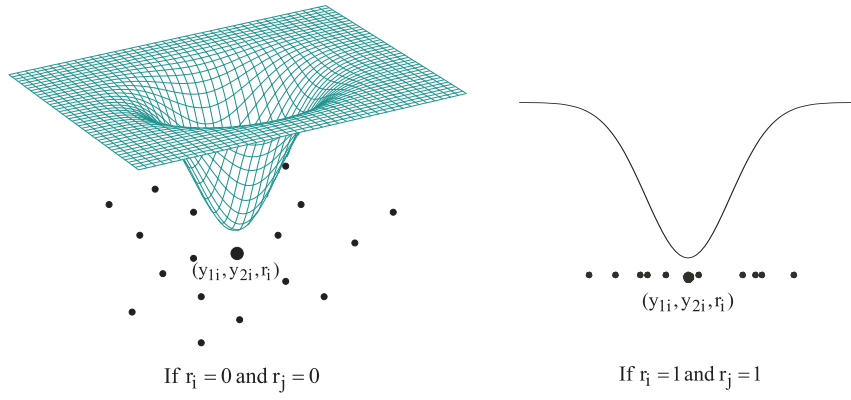


Fig. 5. Shape of the weights. On the left hand side the weights are shown for the situation  $r_i = 0$  and  $r_j = 0$  (completers), while on the right hand side the weights are shown for the situation  $r_i = r_j = 1$  (incompleters).

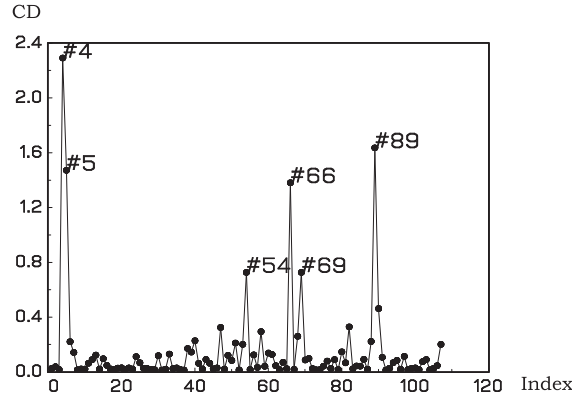


Fig. 6. Influential subjects of the mastitis data for the kernel weighted global influence with initial bandwidths  $\tilde{g}_1 = \tilde{g}_2 = 0.2$ .

consideration. We suggest to use a density adaptive bandwidth  $g = g_1 = g_2$ . Let  $(y_{1i}, y_{2i}, r_i)$  be the observation of interest. If  $r_i = 0$ , the bandwidth  $g$  is taken as

$$g(y_{1i}, y_{2i}, r_i) = \frac{CK^2(0)}{\sum_{j, r_j=0} K\left(\frac{y_{1j}-y_{1i}}{\tilde{g}_1}\right)K\left(\frac{y_{2j}-y_{2i}}{\tilde{g}_2}\right)}. \quad (13)$$

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

$$g(y_{1i}, y_{2i}, r_i) = \frac{CK^2(0)}{\sum_{j, r_j=1} K\left(\frac{y_{1j}-y_{1i}}{\tilde{g}_1}\right)K(0)}, \quad (14)$$

where  $C$  is a constant and  $\tilde{g}_1$  and  $\tilde{g}_2$  are two initially chosen bandwidths. Throughout the paper we used the standard normal density function as the kernel function  $K$ .

A kernel weighted global influence analysis with initial bandwidths  $\tilde{g}_1 = \tilde{g}_2 =$

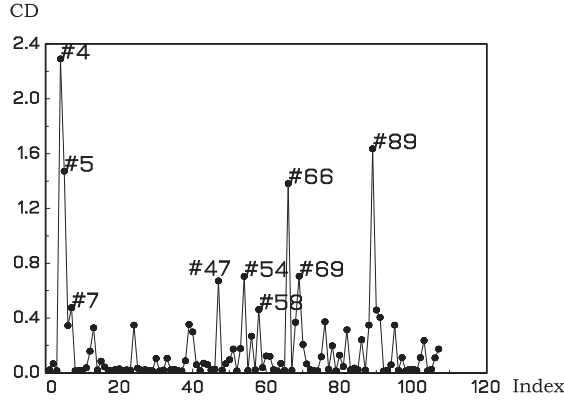


Fig. 7. Influential subjects of the mastitis data for the kernel weighted global influence with initial bandwidths  $\tilde{g}_1 = \tilde{g}_2 = 1.5$ .

0.2 and  $\tilde{g}_1 = \tilde{g}_2 = 1.5$  on the mastitis data leads to Figures 6 and 7 respectively. 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 ML estimators  $\hat{\gamma}_{(-54)}$ ,  $\hat{\gamma}_{(-69)}$  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. If we have a closer look at Figure 7, a second group 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 also be seen as outlying observations with substantial influence. A comparison of Figures 6 and 7 in this respect clearly shows the role of the bandwidth as a tuning parameter in an explorative sensitivity analysis. Both figures show the same influential complete cases but Figure 7 with the larger bandwidth adds to these some incomplete influential cases.

#### 4.2 Kernel Weighted Local Influence

The local influence approach can be extended by looking at the direction determined by the neighborhood  $N(i)$ . First, note that from the discussion in Section 3 it follows that  $\mathbf{h}_i = (1 - \mathbf{w}_{(-i)})/D$  where  $D$  is a normalizing constant such that  $\mathbf{h}_i$  has unit length. This motivates the definition of the kernel weighted local influence of the  $i$ th observation  $(y_{1i}, y_{2i}, r_i)$  as

$$C_{\mathbf{h}_{N(i)}} = 2 \left| \mathbf{h}_{N(i)}' \Delta' \ddot{\mathbf{L}}^{-1} \Delta \mathbf{h}_{N(i)} \right|, \quad (15)$$

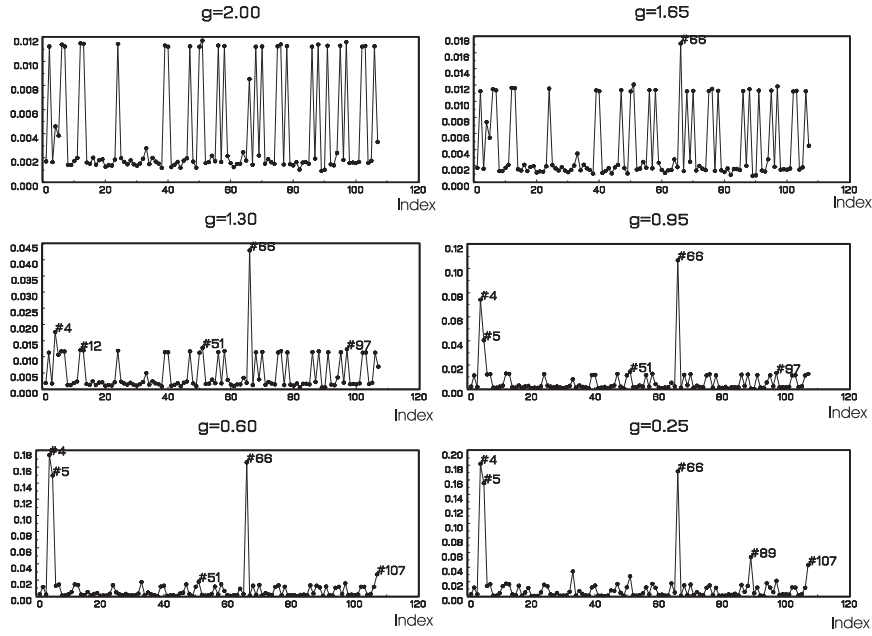


Fig. 8. Influential subjects of the mastitis data for the kernel weighted local influence (increments) with different bandwidths  $g = g_1 = g_2$ .

where

$$\mathbf{h}_{N(i)} = (1 - \mathbf{w}_{(-N(i))})/D, \quad (16)$$

with  $\mathbf{w}_{(-N(i))}$  as defined in (11) and  $D$  a normalizing constant. The choice  $\mathbf{h}_{N(i)}$  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.

It is again interesting to compute the kernel weighted local influence for a series of bandwidths. Because the vector  $\mathbf{h}_{N(i)}$  is normalized, there is no need to have a density-adaptive bandwidth as in Section 4.1.

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.

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) and, if they do not dropout, they are highly influential. This again illustrates the usefulness to examine the kernel weighted influence measures over a range of bandwidth values. 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  $(y_{1i}, y_{2i}, r_i)$ ,  $i = 1, \dots, N$ , we now consider weights centered at points  $(y_1, y_2, r)$  on a one- or two-dimensional grid (for  $r = 1$  and  $r = 0$  respectively) enclosing the full observed data range. Define, in analogy with definition (11), the kernel based weight vector  $\mathbf{w}_{(-N(y_1, y_2, r))}$  as the vector of length  $N$  with elements, for  $j = 1, \dots, N$ ,

$$(\mathbf{w}_{(-N(y_1, y_2, r))})_j = \left[ K(0) \{K(0)\}^{(1-r)} - K\left(\frac{y_{1j} - y_1}{g_1}\right) \{K\left(\frac{y_{2j} - y_2}{g_2}\right)\}^{(1-r)} I(r_j = r) \right] / D, \quad (17)$$

where, as before,  $D$  is a normalization constant such that  $\sum_{i=1}^N (\mathbf{w}_{(-N(y_1, y_2, r))})_j = N$ , and define the kernel weighted global influence measure on the grid points  $(y_1, y_2, r)$  as

$$CD_{N(y_1, y_2, r)} = CD(\mathbf{w}_{(-N(y_1, y_2, r))}). \quad (18)$$

Examining the graph of  $CD_{N(y_1, y_2, r)}$  as a function of  $y_1$  (incompleters) or  $y_1$  and  $y_2$  (completers) allows us to identify influential regions over a grid, not only centered at the observed data points.

The kernel weighted local influence can be calculated over a grid in a similar way. With  $\mathbf{h}_{N(y_1, y_2, r)} = (1 - \mathbf{w}_{(-N(y_1, y_2, r))})/D$  ( $D$  a normalizing constant), define the grid based weighted local influence as  $C\mathbf{h}_{N(y_1, y_2, r)}$ . A plot of the weighted local influence values can be constructed and can lead to additional insights.

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 4.1, we used



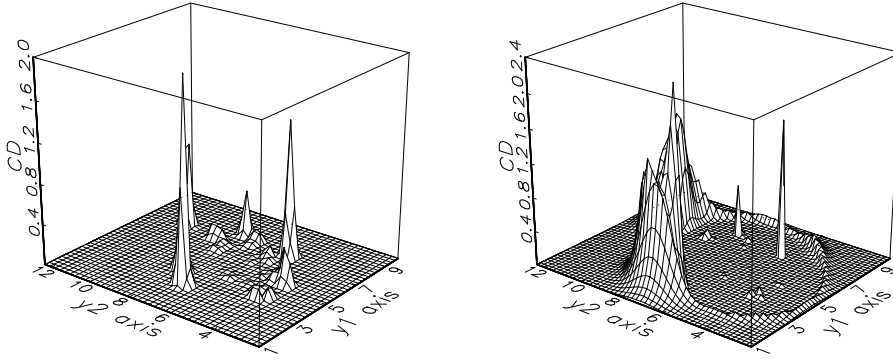


Fig. 9. Kernel weighted global influence graph over a grid of completers with density-adaptive bandwidths initially equal to 0.2 (upper panel) and 1.5 (lower panel).

a density-adaptive bandwidth. The initial bandwidths  $\tilde{g}_1$  and  $\tilde{g}_2$  in (13) and (14) were chosen equal to 0.2 and 1.5 respectively.

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 instead of missing at random. The peaks shown in Figure 9 confirm the results from Section 4.1. 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 types of observations.

The construction of such a grid-based global influence graph is very computer intensive due to the calculation of the numerous (weighted) ML estimates. This is not the case for a grid analysis based on kernel weighted local influence, which is computationally much simpler. So, for the local influence measures, based on the directions  $\mathbf{h}_{N(y_1, y_2, r)}$ , we used a wider range, a finer grid and tried several bandwidth choices. Figure 10 shows a selection of weighted local influence graphs, for four different bandwidths. 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

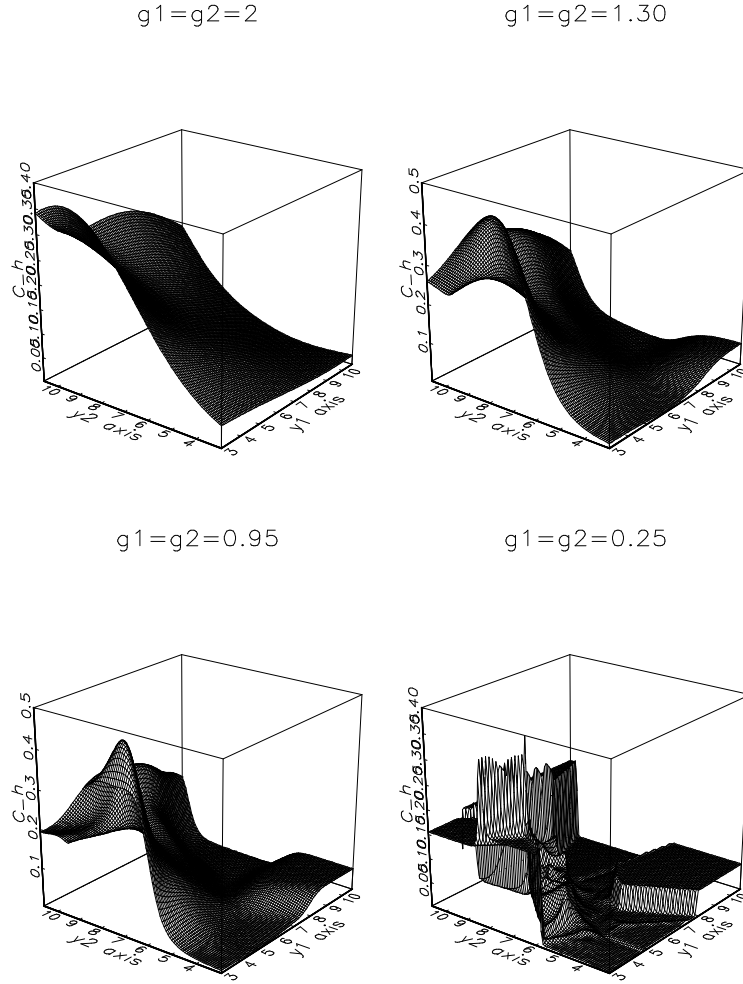


Fig. 10. Kernel weighted local influence graphs over a grid of completers for several bandwidths  $g_1 = g_2$ .

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 at 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 4.2.

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  $(3.5, 7.5)$ .

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

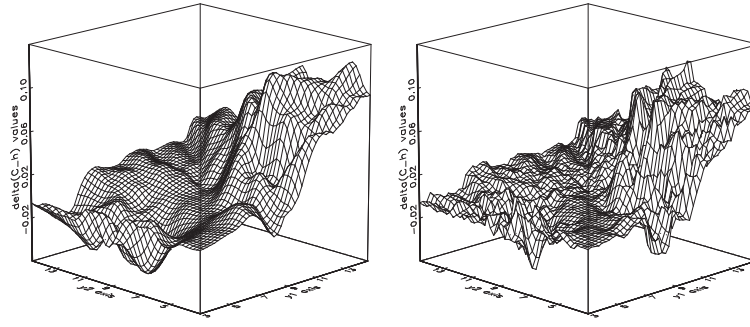


Fig. 11. A figure of the relative average gain in influence of the completers when generating 3 subjects under MNAR. The bandwidths used are respectively equal to 1 and 0.5.

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:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N} \left[ \begin{pmatrix} 6.426 \\ 7.095 \end{pmatrix}, \begin{pmatrix} 2.865 & 2.324 \\ 2.324 & 2.865 \end{pmatrix} \right]. \quad (19)$$

The dropout process was generated according to the following model

$$\text{logit}[P(R_i = 1|Y_{i1}, Y_{i2})] = -3.379 + 0.387Y_{i1} + \psi_2 Y_{i2}, \quad (20)$$

where  $\psi_2$  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.

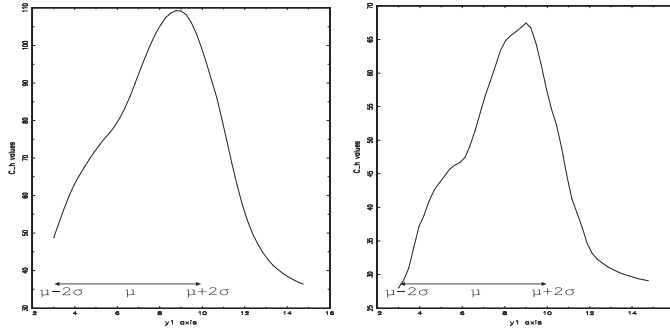


Fig. 12. A figure of the relative average gain in influence of the incompleters when generating 3 subjects under MNAR. The bandwidths used are respectively equal to 1 and 0.5.  $\mu$  and  $\sigma$  denote the mean and standard deviation of  $Y_1$ .

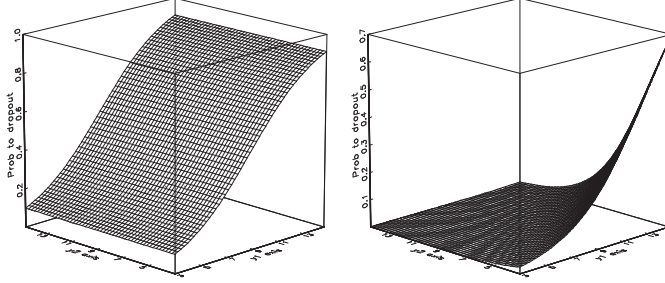


Fig. 13. Plot of the probability of dropout. On the left hand side the dropout probability under MAR is shown, while on the right hand side the dropout probability is shown under MNAR.

### 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  $\psi_2$  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.

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

For the types of observations with a missing second measurement the largest

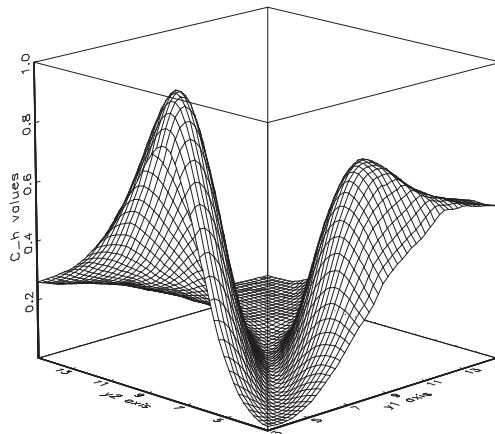


Fig. 14. The average kernel weighted local influence for the completers of the 100 reference datasets

influence is located at higher  $y_1$  values as can be seen in Figure 12. A high value of  $y_1$  often goes with a higher value of  $y_2$  (correlation 0.8), a combination which has, under the MNAR model, a small probability to drop out. If it then drops out nevertheless, it is highly influential.

## 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. 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.

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. The larger influence of observations with a high  $Y_1$ -measurement and a low  $Y_2$ -measurement is not surprising. In Figure 16 a scatterplot of the completers is given. If we consider the

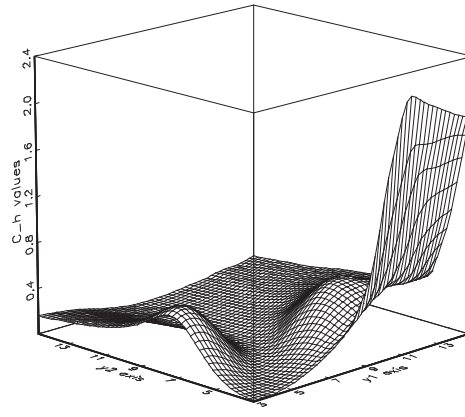


Fig. 15. Kernel weighted local influence for the completers of the 100 complete datasets with MNAR dropouts for  $Y_2 > 8.5$

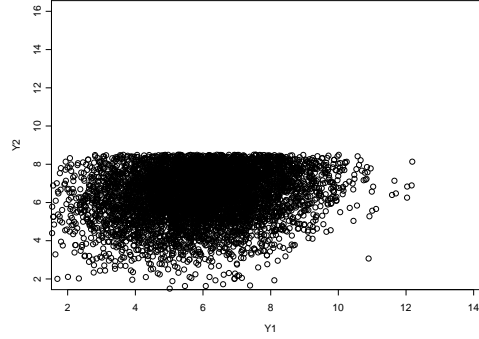


Fig. 16. A scatterplot for all simulated datasets with MNAR dropouts for  $Y_2 > 8.5$

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 17 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.

Other simulation settings (such as larger sample sizes) confirm these results.

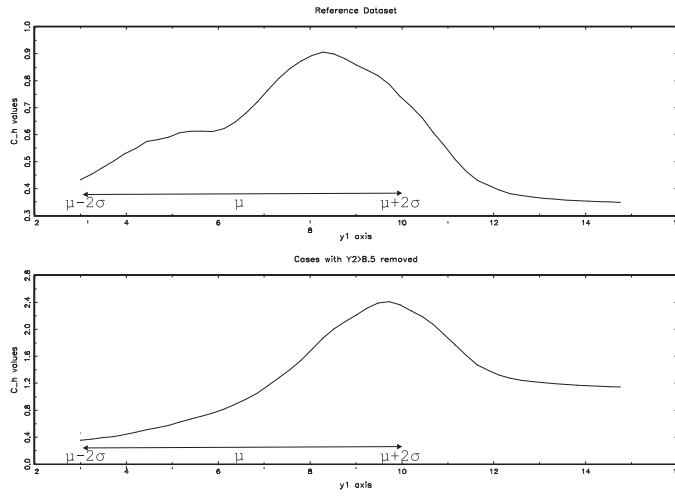


Fig. 17. The figures of kernel weighted local influence for the incompleters of the complete dataset and the incompleters of the datasets with MNAR dropouts for  $Y_2 > 8.5$

The main idea is illustrated here and therefore these other simulations are omitted from this paper.

## 7 Final Remarks

In this paper we introduced some new exploratory and graphical techniques, supplementing existing tools for sensitivity analysis. These methods combine parametric global and local influence measures with nonparametric smoothing weights. They provide new insights in the influence of certain types of observations and offer a nice solution to the problem of masking. The discussion here has been focusing on the setting of two (repeated) measurements. In case of three or more measurements, the kernel based weights (11) and (21) can be based on higher dimensional kernels. Alternatively, one can first determine the Euclidean distance between two observations (belonging to the same pattern) in combination with a one-dimensional kernel function. This latter option leads to the following extension of the weights (11), to any number of measurements.

Let  $(\mathbf{y}_i, \mathbf{r}_i)$  denote the data where  $\mathbf{y}_i = (y_{i1}, \dots, y_{in}) = (\mathbf{y}_i^o, \mathbf{y}_i^m)$  is the vector of observed components  $\mathbf{y}_i^o$  and missing components  $\mathbf{y}_i^m$  and where  $\mathbf{r}_i = (r_{i1}, \dots, r_{in})$  is the vector grouping the missingness indicators

$$r_{il} = \begin{cases} 1 & \text{if } y_{il} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

For a neighborhood  $N(i)$  of outcome  $(\mathbf{y}_i, \mathbf{r}_i)$ , define the weights

$$(\mathbf{w}_{(-N(i))})_j = \{K(0) - K(\|\mathbf{y}_j^o - \mathbf{y}_i^o\|/g)I(\mathbf{r}_j = \mathbf{r}_i)\}/D, \quad (21)$$

where  $K$  is for instance a Gaussian kernel function,  $g$  is the bandwidth and  $D$  a normalizing constant, as before. So, similar to the weights (11), the weights (21) are constant for all observations with a different missingness pattern ( $\mathbf{r}_j \neq \mathbf{r}_i$ ) and assign low weights to all observations  $\mathbf{y}_j$  in the close neighborhood of  $\mathbf{y}_i$  and with an identical missingness pattern ( $\mathbf{r}_j = \mathbf{r}_i$ ). Note that this definition is not restricted to monotone dropout missingness mechanisms.

As a further generalization one could extend the concept of the neighborhood of a particular observation  $(\mathbf{y}_i, \mathbf{r}_i)$  to all observations with not only an identical missingness pattern  $\mathbf{r}_i$  but also with a similar pattern, in this way including, for example, observations which dropout one time point earlier or later. This could be an interesting option in order to enlarge the number of effective observations in the neighborhood of  $(\mathbf{y}_i, \mathbf{r}_i)$  which is, especially in case of several measurements and in view of the curse of dimensionality, not unimportant.

A deeper study of the properties and the applicability of this extension to more than two measurements is beyond the scope of this paper. It is the subject of current and future research.

## Acknowledgments

We wish to thank the referee and the associate editor for valuable remarks leading to an improved presentation. We gratefully acknowledge support from 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 non-



- random 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.
- Ryan, T.P.(1997) *Modern Regression Methods*. New York: Wiley.
- Thijs, H., Molenberghs, G. and Verbeke, G.(2000) The Milk Protein Trial: Influence analysis of the Dropout Process. *Biometrical Journal*, 42, 1–30.
- Thijs, H., Molenberghs, G. and Verbeke, 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.
- Wand, M.P. and Jones, M.C.(1995) *Kernel Smoothing*. London: Chapman & Hall.