Simple consistent cluster methods based on redescending M-estimators with an application to edge identification in images*

by Christine H. Müller and Tim Garlipp
Carl von Ossietzky University of Oldenburg

August 28, 2003

Abstract

We use the local maxima of a redescending M-estimator to identify cluster, a method proposed already by Morgenthaler (1990) for finding regression clusters. We work out the method not only for classical regression but also for orthogonal regression and multivariate location and show that all three approaches are special cases of a general approach which includes also other cluster problems. For the general case we show consistency for an asymptotic objective function which generalizes the density in the multivariate case. The approach of orthogonal regression is applied to the identification of edges in noisy images.

Keywords: Kernel density estimation, M-estimation, consistency, multivariate cluster, regression cluster, orthogonal regression, edge identification in noisy images.
AMS Subject classification: 62 H 30, 62 G 35, 62 G 07, 62 J 05, 62 P 99

*Research supported by the grant Mu 1031/4-1 of the Deutsche Forschungsgemeinschaft.

1 Introduction

Consider a data set with a possibility of different clusters (or subgroups or substructures). One of the important problems of statistical analysis is to identify all these clusters. For example, in image analysis, each cluster may correspond to a different geometric primitive and finding each geometric primitive is therefore an important step towards identifying and locating an object in the environment. For identifying different regression lines, Späth (1979) was one of the first who proposed an algorithm. This cluster method is based on the least squares method. Morgenthaler (1990) and Meer and Tyler (1998) proposed to use M-estimators with redescending score functions to detect different regression clusters. Usually the redescending M-estimators have the disadvantage that the objective function has several local maxima (or minima, respectively). But for identifying substructures in the data this is an advantage since each local maximum may correspond to a substructure. This advantage of redescending M-estimators was also used by Hennig (1997, 2000, 2003) but in a different approach. He used instead of the usual equation for defining M-estimators a fix point version and estimates simultaneously the regression and scale parameter with an indicator function as score function. The M-estimator with indicator function and fixed scale provides the well known method of Hough transform in computer vision (e.g. Stewart 1997). In this paper we follow more the original approach of Morgenthaler and use in particular redescending score functions which satisfy some smoothness conditions. It is known that M-estimators with a smooth score function have many convenient statistical properties like Fréchet differentiability (see e.g. Huber 1981, p.37, or Bednarski et al. 1991).

In the original approach of Morgenthaler, there is the problem that some of the local maxima have no relation to a useful substructure. For example, a local maximum in clusterwise regression can correspond to a line which is orthogonal to the line of a regression cluster. Therefore, Chen et al. (2001) and Arslan (2002) proposed rather complicated additional measures for identifying the local maxima which correspond to real regression cluster. In this paper we show that the problem of identifying the right local maxima is a problem of too few data and a too large scale parameter. Using a too large scale parame-

ter means usually that the redescending M-estimators behave more like the least squares estimators so that the largest (or only) local maximum correspond to a fit over all data and not to a fit of a subgroup. This also happens if the scale is simultaneously estimated (see Arslan 2002). This can be avoided by using a scale parameter which converges to zero with growing sample size. For this situation, we derive here an asymptotic objective function which is independent of the score function. We demonstrate that the largest local maxima of this asymptotic objective function correspond to the largest regression clusters while small local maxima of this objective function correspond to small clusters or provide features which have no relation to clusters. Moreover, we show consistency to the local maxima of the asymptotic objective function so that the asymptotic behavior should hold also approximately for the finite sample case.

While there exists a huge number of different cluster methods, consistency results for them are rather rare. This is caused by the fact that cluster methods belong more to data mining methods so that it is of less interest what are the true cluster and cluster center points. But once true cluster and cluster center points are given, the question about consistency is important. For cluster methods for multivariate data, consistency results are proved as that of Pollard (1981) for the K-means methods and of Davies (1988) for mixtures of elliptical distributions. Also for clusterwise regression, consistency results exists as those of Kiefer (1978), Desarbo and Cron (1988) and Hennig (1997, 2001, 2003). Here we show the consistency for a rather general class of cluster methods which are not restricted to the regression case and may be used for identifying other geometric primitives.

We apply the cluster method based on redescending M-estimators to the problem of identifying edges in noisy images. Cluster methods and redescending M-estimators are widely used in computer vision (e.g. Krishnapuram and Freg 1992, Roth and Levine 1993, Stewart 1997, Müller 2002). All these approaches work with the pixel values themselves. Here we apply the cluster method to points which are calculated from the pixel values by a method proposed by Qiu (1997). The method of Qiu provides points which are distributed around the true edges so that finding the regression clusters with redescending M-estimators leads to very accurate edge estimates. Moreover, the method of Qiu provides

a value for the scale parameter and good starting points for finding the local maxima of the objective function.

We start in Section 2 with the problem of clustering multivariate data coming from a mixture of densities. A main feature of our consistency result is that redescending M-estimators can be regarded as kernel density estimators where the scale parameter is the bandwidth, a connection which recently was used also by Chen and Meer (2002) in computer vision. For multivariate data, it is well known that, letting the scale parameter (bandwidth) go to zero, the objective function converges to the density (see e.g. Scott, 1992). Hence Section 2 repeats more or less known results. However, this section is important for motivating the approaches for classical (vertical) regression in Section 3, orthogonal regression in Section 4 and the general case in Section 5. In particular our asymptotic objective function is a generalization of the density in the multivariate case.

It turns out in Section 3, that in the case of two regression clusters the asymptotic objective function for classical (vertical) regression has largest local maxima at parameters corresponding to the regression clusters. Other local maxima are significantly smaller if they exist. For orthogonal regression the same result is shown in Section 4. We also show in Section 3 that the classical regression, besides its dependence on the choice of the axis and the special error structure, has the disadvantage that the consistency result does not hold for discrete distributions of the regressors. All these disadvantages disappear for orthogonal regression. This is the reason why we use orthogonal regression for identifying edges in noisy images in Section 6. But before presenting the application in image analysis, we show in Section 5 that the cluster approaches for multivariate location, classical regression and orthogonal regression are special cases of a general approach which includes also other cluster problems as those in multivariate regression. For this general approach, we present in Section 5 the consistency result in detail. All proofs are given in the appendix.

2 Clustering of multivariate data

Let $y_{1N}, \ldots, y_{NN} \in \mathbb{R}^k$ be the realization of independent and identically distributed k-dimensional random vectors $Y_{1N}, \ldots, Y_{NN}$, h the density function of the distribution of Y_{nN} for $n = 1, \ldots, N$, and $y_N = (y_{1N}, \ldots, y_{NN})$. The aim is to find the positions of the local maxima of h. These positions of the local maxima are considered as the true center points of the true clusters and the set of these center points is denoted by $\mathcal{M}$, i.e.

$$\mathcal{M} := \{\mu \in \mathbb{R}^k;\ h(\cdot) \text{ has local maximum at } \mu\}.$$

Via the center points the true clusters are defined as the sets of those points which are closest to the center point. I.e. the true cluster with respect to a center point $\mu_0 \in \mathcal{M}$ are those points $y \in \mathbb{R}^k$ so that $\mu_0 \in \arg\min\{\|y - \mu\|;\ \mu \in \mathcal{M}\}$.

In particular, if the distribution of Y_{nN} is a mixture of distributions with unimodal densities, then the density of the distribution h of Y_{nN} has several local maxima. For example, if $Y_{nN} = \mu_l + E_{nN}$ with probability γ_l and E_{nN} has density f_l with maximum at 0, then h is the mixture of densities f_l given by

$$h(\mu) = \sum_{l=1}^{L} \gamma_l f_l(\mu - \mu_l)$$

with $\gamma_l > 0$ and $\sum_{l=1}^{L} \gamma_l = 1$. If the functions $f_l(\cdot - \mu_l)$ have supports which are not overlapping then the local maxima of h are attained at the μ_l. However, if the supports are overlapping as it is the case for multivariate normal distributions, then the local maxima of h do not coincide with the modes of the $f_l(\cdot - \mu_l)$ but they are closely related if the overlap is not too large. Since in practise the densities f_l of the mixture distribution are not known it is better to define center points and clusters not via the modes of the distributions $f_l(\cdot - \mu_l)$. The above definition of center points and clusters is more appropriate for situations with a general density h. Hence the main aim is to estimate the positions of the local maxima of h.

Having the result that kernel density estimates are consistent estimates of the density h (see e.g. Silverman 1986), consistent estimates for the local maxima of h and thus the center points can be defined as the local maxima of the estimated density given by the

kernel estimator. This provides automatically also consistent estimates of the clusters. However, for the consistency of the local maxima, technical problems must be solved.

A kernel density estimator for $h(\mu)$ is given by

$$H_N(\mu, y_N) := \frac{1}{N} \sum_{n=1}^{N} \frac{1}{s_N^k} \rho \left(\frac{1}{s_N} (y_{nN} - \mu) \right),$$

where $\mu \in I\!R^k$, $\rho : I\!R^k \to I\!R^+$ is the kernel function and $s_N \in I\!R^+ \setminus \{0\}$ is the bandwidth. If s_N converges to zero when N tends to infinity, then $H_N(\mu, y_N)$ converges to $h(\mu)$ in probability under some regularity conditions. Let

$$\mathcal{M}_N(y_N) := \{\mu \in I\!R^k;\ H_N(\cdot, y_N) \text{ has local maximum at } \mu\},$$

where the local maxima of $H_N(\cdot, y_N)$ can be found by Newton Raphson method starting at any y_{nN} with $n = 1, \ldots, N$.

Then $\mathcal{M}_N(y_N)$ is the estimate of the set $\mathcal{M}$ of the positions of the true local maxima. The set $\mathcal{M}_N(y_N)$ can be also considered as M-estimates with respect to the objective function $H_N(\cdot, y_N)$. Then ρ is the score function of the M-estimator and s_N is a scale parameter.

Usually ρ will be an unimodal density like that of the standard normal distribution, where $\rho(y) = \frac{1}{(\sqrt{2\pi})^k} e^{-\frac{1}{2} y^\top y}$. Hence each y_{nN} is a candidate for a local maximum of $H_N(\mu, y_N)$. If the bandwidth (scale parameter) s_N is small enough and the distance between the y_{nN} are large enough then these candidates can be really local maxima. But usually there is so much overlap of the $\rho \left(\frac{1}{s_N} (y_{nN} - \mu) \right)$ that no one of the y_{nN} is a local maxima. However, searching the local maxima in increasing direction starting at any y_{nN} should provide the relevant maxima. This is an approach used also by Chu et al. (1998) for constructing corner preserving M-smoother for image reconstruction. The consistency of these M-smoothers even at jumps was shown by Hillebrand and Müller (2001). A similar proof can be used here to prove the consistency of the set $\mathcal{M}_N(y_N)$.

In Section 5 it is shown that the set $\mathcal{M}_N(y_N)$ is a consistent estimator for the set $\mathcal{M}$. For that we need not only pointwise convergence of $H_N(\mu, y_N)$ to $h(\mu)$ but also uniform convergence which can be achieved by intersecting $\mathcal{M}_N(y_N)$ with a compact subset of $I\!R^k$.

Appropriate compact subsets are given by

$$\Theta_\eta := \left\{ \mu \in I\!R^k;\ h(\mu) \geq \frac{1}{\eta} \right\} \text{ with } \eta \in I\!N.$$

3 Clustering of regression data

In the classical regression model with one cluster, we have independent and identically distributed observations $Z_{1N} := (X_{1N}^\top, Y_{1N})^\top, \ldots, Z_{NN} := (X_{NN}^\top, Y_{NN})^\top$ with

$$Y_{nN} = \left(1, X_{nN}^\top\right) \beta_0 + E_{nN}$$

for $n = 1, \ldots, N$, where X_{nN} and E_{nN} are independent, $\beta_0 \in I\!R^p$, $p \geq 2$, is an unknown parameter vector and Y_{nN} and X_{nN} are observed. Let f denote the density function of the distribution of E_{nN} and G_0 the distribution of X_{nN}. Moreover, let $z_N = (z_{1N}, \ldots, z_{NN}) = (x_{1N}^\top, y_{1N})^\top, \ldots, (x_{NN}^\top, y_{NN})^\top)$ a realization of $Z_N = (Z_{1N}, \ldots, Z_{NN})$.

The M-estimator for β is defined as a maximum point of the objective function

$$H_N(\beta, z_N) := \frac{1}{N} \sum_{n=1}^{N} \frac{1}{s_N} \rho \left(\frac{1}{s_N} \left(y_{nN} - \left(1, x_{nN}^\top\right) \beta \right) \right),$$

where $\rho : I\!R \to I\!R^+$ is the score function and $s_N \in I\!R^+ \setminus \{0\}$ is a scale parameter (see e.g. Huber 1973, 1981, Hampel et al. 1986). If ρ is not convex, that is the derivative of ρ is redescending, then $H_N(\cdot, z_N)$ has several local maxima so that we define

$$\mathcal{M}_N(z_N) := \{\beta \in I\!R^p;\ H_N(\cdot, z_N) \text{ has local maximum at } \beta\}. \tag{1}$$

The local maxima of $H_N(\cdot, z_N)$ can be found by Newton Raphson method starting at any hyperplane through $(x_{n_1 N}^\top, y_{n_1 N})^\top, \ldots, (x_{n_p N}^\top, y_{n_p N})^\top$ with $\{n_1, \ldots, n_p\} \subset \{1, \ldots, N\}$.

It is shown in Section 5 that under some regularity conditions for $s_N \to 0$, we have

$$H_N(\beta, Z_N)) \stackrel{N \to \infty}{\longrightarrow} h(\beta) := \int f\left(\left(1, x^\top\right) (\beta - \beta_0) \right) G_0(dx)$$

in probability for all $\beta \in I\!R^p$. The function h plays now the same role as the density h in multivariate density estimation. In particular it is independent of the score function

ρ. It would depend on ρ if the scale parameter would not converge to zero, which is in particular the case if the scale is simultaneously estimated as in Arslan (2002). Moreover, we have the following Fisher consistency result.

Lemma 1 *If f is a unimodal density with maximum at 0, then h has only one maximum at β_0.*

Now regard the situation of L regression clusters with different parameter vectors β_l, $l = 1, \ldots, L$. Then the n'th observation is given by

$$Y_{nN} = \left(1, X_{nN}^{\top}\right) \beta_l + E_{nN}$$

if it is coming from the l'th cluster. Since the distribution of the regressors X_{nN} may also depend on the cluster, we use G_l for the distribution of X_{nN} coming from the l'th cluster. In Section 5 it is shown that for $s_N \to 0$ we now have

$$H_N(\beta, Z_N) \overset{N\to\infty}{\longrightarrow} h(\beta) := \sum_{l=1}^{L} \gamma_l \int f\left(\left(1, x^{\top}\right)(\beta - \beta_l)\right) G_l(dx)$$

in probability for all $\beta \in I\!R^p$. Again $\gamma_l > 0$ denotes the probability that the n'th observation is coming from the l'th cluster and it holds $\sum_{l=1}^{L} \gamma_l = 1$.

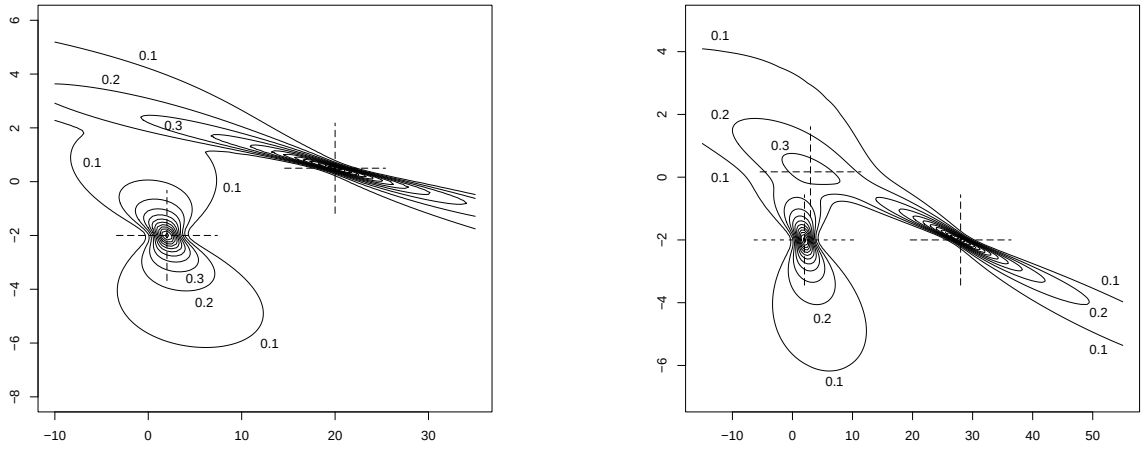

Figure 1: Contour plots of the limit function h in the case of two nonparallel lines (left, $\beta_1 = (2, -2)$, $\beta_2 = (20, 0.5)$, $\gamma_1 = \gamma_2 = 0.5$, $G_1 \sim \mathcal{N}(1, 3)$, $G_2 \sim \mathcal{N}(10, 3)$, $f = f_{\mathcal{N}(0,1)}$) and the case of two parallel lines (right, $\beta_1 = (2, -2)$, $\beta_2 = (28, -2)$, $\gamma_1 = \gamma_2 = 0.5$, $G_1 \sim \mathcal{N}(1, 3)$, $G_2 \sim \mathcal{N}(10, 3)$, $f = f_{\mathcal{N}(0,1)}$)

Under enough separation the local maxima of h are attained at $\beta_1, \ldots, \beta_L$. This is analogous to the multivariate case. Enough separation means here that the density f has a compact support and that the parameter vectors β_l and the supports of the distributions G_l are enough separated. If the separation is not enough then at least the local maxima are attained close to $\beta_1, \ldots, \beta_L$. This can be seen also from the left contour plot of h in Figure 1 in the case of two regression clusters given by two nonparallel lines, where the regressors and the errors have normal distributions, namely $G_1 \sim \mathcal{N}(1, 3)$, $G_2 \sim \mathcal{N}(10, 3)$, $f = f_{\mathcal{N}(0,1)}$. However, in the case of two parallel lines as given in the right plot in Figure 1, there appears a third local maximum which correspond to a line more or less orthogonal to the parallel lines depending on the distribution of the regressors. But the height of this third local maximum is much smaller than the height of the other local maxima so that it can be easily separated from the other local maxima. See also Section 6.

Hence as in the multivariate case we will regard the positions of the local maxima of h as the true parameter vectors which shall be estimated. Let $\mathcal{M}$ be the set of the positions of these local maxima, i.e.

$$\mathcal{M} := \{\beta \in I\!R^p;\ h(\cdot) \text{ has local maximum at } \beta\}.$$

The regression hyperplanes $l(\beta) := \left\{(x^\top, y)^\top \in I\!R^p;\ y = \left(1, x^\top\right)\beta\right\}$ given by maximum points $\beta \in \mathcal{M}$ are the true center points (center planes) of the regression clusters. The cluster belonging to a center plane given by β_l are all $(x^\top, y)^\top \in I\!R^p$ for which $l(\beta_l)$ is the closest plane, i.e. all $(x^\top, y)$ such that $\beta_l \in \arg\min\left\{|y - \left(1, x^\top\right)\beta|;\ \beta \in \mathcal{M}\right\}$. Hence, if we can estimate $\mathcal{M}$ consistently, then also the regression clusters are consistently estimated. If some of the local maxima of $\mathcal{M}$ are not related to real clusters as the third local maxima in the example of two parallel lines they can be excluded afterwards by the height of the maximum or, another possibility, by the size of the corresponding cluster. But the first step is to estimate consistently all local maxima of h.

The estimate for $\mathcal{M}$ is the set $\mathcal{M}_N(z_N)$ defined by (1). As for the multivariate case, the consistency of $\mathcal{M}_N(z_N)$ can be shown in Section 5 only if $\mathcal{M}_N(z_N)$ is intersected with a compact set which is here

$$\Theta_\eta := \left\{\beta \in I\!R^p;\ h(\beta) \geq \frac{1}{\eta}\right\} \text{ with } \eta \in I\!N.$$

However, here is the compactness of Θ_η not always satisfied. In particular, as the following lemma shows, it is not satisfied if one of the distributions G_l is discrete so that regression experiments with repetitions at finite design points are excluded.

Lemma 2 *If f is continuous then Θ_η is compact for all $\eta \in \mathbb{N}$ if and only if none of distributions G_l has positive mass on a subspace of $\mathbb{R}^{p-1}$.*

4 Clustering of regression data by orthogonal regression

For orthogonal regression usually an error-in-variable model is assumed. This means that, in the case of one cluster, we have independent and identically distributed observations $Z_{1N} := (V_{1N}^\top, W_{1N})^\top, \ldots, Z_{NN} := (V_{NN}^\top, W_{NN})^\top$ with

$$(V_{nN}^\top, W_{nN}) = (X_{nN}^\top, Y_{nN}) + (E_{1nN}^\top, E_{2nN})$$

for $n = 1, \ldots, N$, where $(X_{nN}^\top, Y_{nN}), E_{1nN}, E_{2nN}$ are independent, X_{nN}, V_{nN}, E_{1nN} are $(p-1)$-dimensional random vectors, Y_{nN}, W_{nN}, E_{2nN} are one-dimensional random variables, and $(X_{nN}^\top, Y_{nN})a_0 = b_0$ almost surely for some unknown, but fixed $(a_0^\top, b_0)^\top \in S_1 \times \mathbb{R}$ where $S_1 = \{a \in \mathbb{R}^p;\ a^\top a = 1\}$. Usually it is assumed that $(E_{1nN}^\top, E_{2nN})^\top$ has a symmetrical, elliptical distribution with density f_0 such that $a^\top (E_{1nN}^\top, E_{2nN})^\top$ has a distribution with density f for all $a \in S_1$. Let $a = (a_1^\top, a_p)^\top$ and $a_0 = (a_{01}^\top, a_{0p})^\top$. W.l.o.g. we can assume that $a_{0p} \neq 0$ so that $Y_{nN} = \frac{b_0}{a_{0p}} - \frac{1}{a_{0p}} a_{01}^\top X_{nN}$, and G_0 denotes the distribution of the regressor X_{nN}. Let again $z_N = (z_{1N}, \ldots, z_{NN}) = \left((v_{1N}^\top, w_{1N})^\top, \ldots, (v_{NN}^\top, w_{NN})^\top\right)$ denote the realization of $Z_N = (Z_{1N}, \ldots, Z_{NN})$.

To avoid working with S_1, we can use a one-to-one mapping $a : C_{p-1} \to S_1$, where C_{p-1} is a compact subset of $\mathbb{R}^{p-1}$. In the two-dimensional case, i.e. $p = 2$, the function a can be given for example by $a(\alpha) = (\cos(\alpha), \sin(\alpha))^\top$ with $\alpha \in [-\pi, \pi] = C_1$.

An M-estimator for (a_0, b_0) was proposed by Zamar (1989) and extends the orthogonal least squares regression estimator. It is defined as a maximum point of the objective

function

$$H_N(a,b,z_N) := \frac{1}{N}\sum_{n=1}^{N}\frac{1}{s_N}\,\rho\left(\frac{1}{s_N}\,(a_1^\top v_{nN} + a_p w_{nN} - b)\right),$$

where $\rho : I\!R \to I\!R^+$ is the score function and $s_N \in I\!R^+ \setminus \{0\}$ is a scale parameter. As for classical vertical regression, if the derivative of ρ is redescending, then $H_N(a,b,z_N)$ has several local maxima so that we define

$$\mathcal{M}_N(z_N) := \{(a^\top, b)^\top \in S_1 \times I\!R;\ H_N(\cdot,\cdot,z_N) \text{ has local maximum at } (a,b)\}. \tag{2}$$

The local maxima of $H_N(\cdot,\cdot,z_N)$ can be found as for vertical regression, i.e. by Newton Raphson method starting at any hyperplane through $(v_{n_1N}^\top, w_{n_1N})^\top, \ldots, (v_{n_pN}^\top, w_{n_pN})^\top$ with $\{n_1,\ldots,n_p\} \subset \{1,\ldots,N\}$.

It is shown in Section 5 that under some regularity conditions for $s_N \to 0$, we have

$$H_N(a,b,Z_N) \overset{N\to\infty}{\longrightarrow} h(a,b) := \int f\left(b - \frac{a_p}{a_{0p}}b_0 - \left(a_1^\top - \frac{a_p}{a_{0p}}a_{01}^\top\right)x\right) G_0(dx)$$

in probability for all $(a^\top, b)^\top \in S_1 \times I\!R$. Again, as in Section 3, h is independent of ρ. Moreover, in opposite to classical vertical regression, the function $h(a,b)$ can be interpreted as a density and shows therefore more relations to the function h in the multivariate case of Section 2. We have namely the following lemma.

Lemma 3 *For every $a \in S_1$, the distribution of $a_1^\top V_{nN} + a_p W_{nN}$ has the density*

$$f_{a_1^\top V_{nN} + a_p W_{nN}}(b) = \int f\left(b - \frac{a_p}{a_{0p}}b_0 - \left(a_1^\top - \frac{a_p}{a_{0p}}a_{01}^\top\right)x\right) G_0(dx).$$

If f is unimodal with maximum at 0, then $f_{a_1^\top X_{nN} + a_p Y_{nN}}(b)$ attains its maximum value at $a = a_0$ and $b = b_0$ so that we have again Fisher consistency in the case of one cluster. This can be seen as in Lemma 1.

Now, we consider a mixture of error-in-variable models. That is $(V_{nN}^\top, W_{nN})^\top$ follows with probability γ_l an error-in-variable model with parameter $(a_l^\top, b_l)^\top = (a_{l1}^\top, a_{lp}, b_l)^\top \in S_1 \times I\!R$ and regressor distribution G_l for $l = 1,\ldots,L$, where $\sum_{l=1}^{L}\gamma_l = 1$. Then the

distribution of $a_1^\top V_{nN} + a_p W_{nN}$ has the density

$$\begin{aligned} h(a,b) &:= f_{a_1^\top V_{nN} + a_p W_{nN}}(b) \\ &= \sum_{l=1}^{L} \gamma_l \int f\left(b - \frac{a_p}{a_{lp}} b_l - \left(a_1^\top - \frac{a_p}{a_{lp}} a_{l1}^\top\right) x\right) G_l(dx) \end{aligned}$$

and $H_N(a, b, Z_N)$ converges to $h(a, b)$ in probability for every $(a^\top, b)^\top \in S_1 \times I\!R$. If the regression hyperplanes given by (a_l, b_l) are enough separated then $h(a, b)$ will have local maxima at $(a, b) = (a_l, b_l)$.

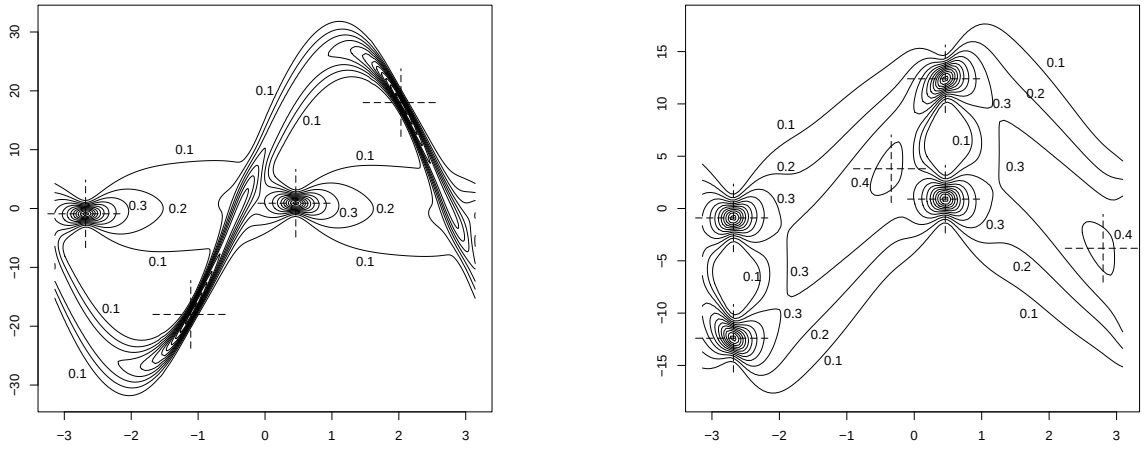

Figure 2: Contour plots of the limit function h in the case of two nonparallel lines (left, $(\alpha_1, b_1) = (0.46, 0.9)$, $(\alpha_2, b_2) = (-1.11, -18)$, $\gamma_1 = \gamma_2 = 0.5$, $G_1 \sim \mathcal{N}(1, 3)$, $G_2 \sim \mathcal{N}(10, 3)$, $f = f_{\mathcal{N}(0,1)}$) and two parallel lines (right, $(\alpha_1, b_1) = (0.46, 0.9)$, $(\alpha_2, b_2) = (0.46, 12.4)$, $\gamma_1 = \gamma_2 = 0.5$, $G_1 \sim \mathcal{N}(1, 3)$, $G_2 \sim \mathcal{N}(10, 3)$, $f = f_{\mathcal{N}(0,1)}$)

See for example Figure 2 for the two-dimensional case with $a_l = (\cos(\alpha_l), \sin(\alpha_l))^\top$. The parameters of the two lines giving the regression clusters are chosen so that the lines are similar to the lines given by the parameters in Figure 1. Therefore also the same distributions for the regressors and the error are chosen although this provides a larger variability of the data since the regressors are also influenced by the error. Nevertheless Figure 2 is rather similar to the Figure 1 taking into account the different parametrization of the lines. Note that the symmetry in Figure 2 is caused by the region $[-\pi, \pi]$ for the parameter α where for example the region $[-\pi, 0]$ would be enough. But from the region $[-\pi, \pi]$ we see better the circular behavior. Hence it turns out for orthogonal regression that, for clusters around nonparallel lines, only two local maxima

appear where, for clusters around two parallel lines, a third local maximum with a rather small height appears. This behavior coincides with the behavior of classical regression treated in Section 3.

The aim is now to estimate the local maxima of $h(a,b)$, or more precisely, the set

$$\mathcal{M} := \{(a^\top, b)^\top \in S_1 \times I\!R;\ h(\cdot,\cdot) \text{ has local maximum at } (a,b)\}.$$

Again, the regression hyperplanes $l(a,b) := \{(v^\top, w)^\top \in I\!R^{p+1};\ a_1^\top v + a_p w = b\}$ given by maximum points $(a^\top, b)^\top \in \mathcal{M}$ are the center points (center planes) of the regression clusters. The cluster belonging to a center plane given by (a_l, b_l) are all $(v^\top, w)^\top \in I\!R^{p+1}$ for which $l(a_l, b_l)$ is the closest plane, i.e. all $(v^\top, w)^\top$ such that $(a_l^\top, b_l)^\top \in \arg\min\{|a_1^\top v + a_p w - b|;\ (a^\top, b)^\top \in \mathcal{M}\}$. If we can estimate $\mathcal{M}$ consistently, then also the regression clusters are consistently estimated.

The estimate for $\mathcal{M}$ is the set $\mathcal{M}_N(z_N)$ defined by (2). As before, the consistency of $\mathcal{M}_N(z_N)$ can be shown in Section 5 only if $\mathcal{M}_N(z_N)$ is intersected with the set

$$\Theta_\eta := \left\{(a^\top, b)^\top \in S_1 \times I\!R;\ h(a,b) \geq \frac{1}{\eta}\right\}.$$

Since a is lying in the compact set S_1 and $h(a,\cdot)$ is a density function the compactness of Θ_η holds here for all distributions G_l. Hence, orthogonal regression is also in this sense superior to classical vertical regression where a restriction on the G_l is necessary to ensure the compactness of Θ_η (see Section 3).

5 Consistency of the center points of multivariate and regression clusters

The approaches for multivariate data and regression data of the Sections 2, 3, and 4 can be combined by regarding

$$H_N(\theta, E_N, X_N) := \frac{1}{N}\sum_{n=1}^{N} \frac{1}{s_N^k}\rho\left(\frac{1}{s_N}\left(E_{nN} - c^n(\theta, X_{nN})\right)\right),$$

where $\theta = (\theta_1, \ldots, \theta_q)^\top \in I\!R^q$, $E_N = (E_{1N}, \ldots, E_{NN})^\top$ is the error matrix (vector), $X_N = (X_{1N}, \ldots, X_{NN})^\top$ is the matrix of regressors, $\rho : I\!R^k \to I\!R^+$, and $c^n(\cdot, x) = (c_1^n(\cdot, x), \ldots, c_k^n(\cdot, x))^\top : I\!R^q \to I\!R^k$ for all $x \in \mathcal{X}$. $E_{1N}, \ldots, E_{NN}, X_{1N}, \ldots, X_{NN}$ are independent with density $f : I\!R^k \to I\!R^+$ for E_{nN} and distribution G_n for X_{nN} on $\mathcal{X}$ for $n = 1, \ldots, N$. We have $G_n = G_l$ and $c^n = c^l$ if the n'th observation is coming from the l'th cluster.

For $i, j \in \{1, \ldots, q\}$, let be

$$
\begin{aligned}
H'_{Ni}(\theta, E_N, X_N) &:= \frac{1}{N} \sum_{n=1}^{N} \frac{1}{s_N^k} \frac{\partial}{\partial \theta_i} \rho \left(\frac{1}{s_N} \left(E_{nN} - c^n(\theta, X_{nN}) \right) \right), \\
H''_{Nij}(\theta, E_N, X_N) &:= \frac{1}{N} \sum_{n=1}^{N} \frac{1}{s_N^k} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \rho \left(\frac{1}{s_N} \left(E_{nN} - c^n(\theta, X_{nN}) \right) \right), \\
h(\theta) &:= \sum_{l=1}^{L} \gamma_l \int f\left(c^l(\theta, x) \right) G_l(dx), \\
h'_i(\theta) &:= \sum_{l=1}^{L} \gamma_l \int \frac{\partial}{\partial \theta_i} f\left(c^l(\theta, x) \right) G_l(dx), \\
h''_{ij}(\theta) &:= \sum_{l=1}^{L} \gamma_l \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} f\left(c^l(\theta, x) \right) G_l(dx),
\end{aligned}
$$

where $\gamma_l > 0$ and $\sum_{l=1}^{L} \gamma_l = 1$.

The multivariate case treated in Section 2 is given with $q = k$, $\theta = \mu$, $\mathcal{X} = \{1\}$, and $c^n(\theta, x) = c^n(\mu, x) = \mu - \mu_n$, so that $H_N(\mu, Y_N)$ of Section 2 is now $H_N(\theta, E_N, X_N)$ where $X_{nN} = 1$ for $n = 1, \ldots, N$. The vertical regression is given with $q = p$, $k = 1$, $\theta = \beta$, $\mathcal{X} \subset I\!R^{p-1}$, and $c^n(\theta, x) = c^n(\beta, x) = \left(1, x^\top\right)(\beta - \beta_n)$, and $H_N(\beta, Z_N)$ of Section 3 is here $H_N(\theta, E_N, X_N)$. For the orthogonal regression, we have $q = p$, $k = 1$, $\theta = (\alpha^\top, b)^\top$, $\mathcal{X} \subset I\!R^{p-1}$, and $c^n(\theta, x) = c^n(\alpha, b, x) = b - \frac{a_p(\alpha)}{a_{np}} b_n - \left(a_1(\alpha)^\top - \frac{a_p(\alpha)}{a_{np}} a_{n1}^\top \right) x$, where $(a_1^\top, a_p)^\top : C_{p-1} \to S_1$ is a one-to-one mapping from the compact set $C_{p-1} \subset I\!R^{p-1}$ onto S_1. The objective function $H_N(a, b, Z_N)$ of Section 4 coincides with $H_N(\theta, E_N, X_N)$ of this section because of the relations $(V_{nN}^\top, W_{nN}) = (X_{nN}^\top, Y_{nN}) + (E_{1nN}^\top, E_{2nN})$ and $Y_{nN} = \frac{b_n}{a_{np}} - \frac{1}{a_{np}} a_{n1}^\top X_{nN}$ and the fact that $a^\top (E_{1nN}^\top, E_{2nN})^\top$ behaves like a E_{nN} with density f. Note that only by using the error matrix E_N we can treat all three cases together. Hence the sets $\mathcal{M}_N(Y_n)$ and $\mathcal{M}_N(Z_N)$ of local maximum points of $H_N(\theta, E_N, X_N)$ of Sections 2, 3, and 4 are here denoted by $\mathcal{M}_N(E_N, X_N)$.

We are going to show that the set $\mathcal{M}_N(E_N, X_N)$ is a consistent estimate of the set $\mathcal{M}$ of local maximum points of h. To show this we have to restrict $\mathcal{M}_N(E_N, X_N)$ to the intersection

$$\mathcal{M}_N^\eta(E_N, X_N) := \mathcal{M}_N(E_N, X_N) \cap \Theta_\eta$$

where for $\eta \in I\!N$

$$\Theta_\eta := \left\{ \theta \in I\!R^q;\ h(\theta) \geq \frac{1}{\eta} \right\}. \tag{3}$$

We will show that $\mathcal{M}_N^\eta(E_N, X_N)$ is a consistent estimate of $\mathcal{M}$ for all η greater a value η_0 so that the intersection provides indeed no restriction.

Under consistency we understand here that $\mathcal{M}_N^\eta(E_N, X_N)$ is lying in a δ neighborhood of $\mathcal{M}$ and vice versa with probability converging to one if N tends to infinity. Hence let

$$\mathcal{U}_\delta(\mathcal{M}_N^\eta(E_N, X_N)) := \{\theta \in I\!R^q;\ \text{ there exists } \theta_0 \in \mathcal{M}_N^\eta(E_N, X_N) \text{ with } \|\theta - \theta_0\| < \delta\}$$

the δ neighborhood of $\mathcal{M}_N^\eta(E_N, X_N)$ and define $\mathcal{U}_\delta(\mathcal{M})$ analoguously. For the consistency we need some assumptions.

For $e = (e_1, \ldots, e_k)^\top$ and $\rho : I\!R^k \to I\!R^+$ let be $\rho'(e) := (\rho'_1(e), \ldots, \rho'_k(e))^\top$ and $\rho''(e) := (\rho''_{rs}(e))_{r,s=1,\ldots,k}$, where $\rho'_r(e) := \frac{\partial}{\partial e_r}\rho(e)$ and $\rho''_{rs}(e) := \frac{\partial}{\partial e_r}\rho'_s(e)$ for $r, s = 1, \ldots, k$. Define $f'(e)$ and $f''(e)$ analogously. Also set $h'(\theta) = (h'_1(\theta), \ldots, h'_q(\theta))^\top$ and $h''(\theta) = (h''_{ij}(\theta))_{i,j=1,\ldots,q}$ and use H'_N and H''_N analogously. Moreover, let $\lambda_{\max} h''(\theta)$ be the maximum eigenvalue of $h''(\theta)$ and $\|A\|$ a norm of the matrix $A = (A_{ij})_{i,j=1,\ldots,q}$, for example $\|A\| = \sqrt{\sum_{i=1}^q \sum_{j=1}^q A_{ij}^2}$. Then we make the following assumptions

[1] $s_N \to 0,\ N\, s_N^{q(k+3)+k+4} \to \infty$ for $N \to \infty$,

[2] The support of G_l is included in $\mathcal{X}$ for all $l = 1, \ldots, L$,

[3] $\frac{\partial^2}{\partial e_r \partial e_s} f(e)$ and $\frac{\partial^2}{\partial e_r \partial e_s}\rho(e)$ are Lipschitz continuous for $r, s = 1, \ldots, k$,

[4] $\frac{\partial^2}{\partial \theta_i \partial \theta_j} c^l(\theta, x)$ is Lipschitz continuous with respect to θ uniformly in $x \in \mathcal{X}$ for $i, j = 1, \ldots, q$ and $l = 1, \ldots, L$,

[5] $\int \rho(v)dv = 1$ and $\int \rho(v)\,||v||dv < \infty$,

[6] $\int \rho''_{rs}(v)^2\,dv < \infty$ and $\int \rho'_r(v)^2\,dv < \infty$ for $r, s = 1, \ldots, k$,

[7] $\sup_{\theta\in\Theta, x\in\mathcal{X}} \left|\frac{\partial}{\partial\theta_i} c^l_r(\theta, x)\right| < \infty$ and $\sup_{\theta\in\Theta, x\in\mathcal{X}} \left|\frac{\partial^2}{\partial\theta_i\,\partial\theta_j} c^l_r(\theta, x)\right| < \infty$ for $r, s = 1, \ldots, k$, $i, j = 1, \ldots, q$, $l = 1, \ldots, L$ and all compact sets $\Theta \subset I\!R^q$.

[8] Θ_η is compact for all $\eta \in I\!N$,

[9] $\mathcal{M}$ is finite and $h''(\theta)$ is negative definite for all $\theta \in \mathcal{M}$,

[10] $\min\{|\lambda_{\max}(h''(\theta))|;\ \theta \in \mathcal{M}_0 \cap \Theta_\eta\} > 0$ for all $\eta \in I\!N$, where $\mathcal{M}_0 := \{\theta \in I\!R^q;\ h'(\theta) = 0 \text{ and } h(\theta) > 0\}$.

Condition [10] is essential but often overlooked in similar approaches (see Hillebrand and Müller 2001). Contrary to the other conditions, it is in general not easy to verify. However, if G_l has a unimodal density and also the density f is unimodal, it is satisfied under enough separation of the clusters. Then we have in particular $\mathcal{M} = \mathcal{M}_0$. The order of the convergence of s_N in Condition [1] can be weakened by using a proof based on Fourier transforms. Since such a proof is more complicated and longer we give here an elementary proof using the stronger assumption.

Under the assumptions we have now the main theorem.

Theorem 1 *Under the conditions [1] - [10], there exists a $\eta_0 \in I\!N$ so that for all $\epsilon > 0$, $\delta > 0$, $\eta \geq \eta_0$ we have: there exists $N_{\eta,\epsilon,\delta} \in I\!N$ with*

$$P\left(\mathcal{M}^\eta_N(E_N, X_N) \subset \mathcal{U}_\delta(\mathcal{M}) \text{ and } \mathcal{M} \subset \mathcal{U}_\delta(\mathcal{M}^\eta_N(E_N, X_N))\right) \geq 1 - \epsilon$$

for all $N \geq N_{\eta,\epsilon,\delta}$.

The proof of the theorem bases on the following three lemmata.

Lemma 4 *Under the conditions [2] - [7], there exists a constant C such that for all $N \in I\!N$, $\theta \in I\!R^q$ and $i, j \in \{1, \ldots, q\}$, we have*

$$|E(H_N(\theta, E_N, X_N)) - h(\theta)| \leq C s_N,$$
$$|E(H'_{Ni}(\theta, E_N, X_N)) - h'_i(\theta)| \leq C s_N,$$
$$|E(H''_{Nij}(\theta, E_N, X_N)) - h''_{ij}(\theta)| \leq C s_N.$$

Lemma 5 *Under the conditions [2] - [7] for all compact sets $\Theta \subset \mathbb{R}^q$, there exists a constant C such that for all $N \in \mathbb{N}$, $\theta \in \Theta$ and $i, j \in \{1, \ldots, q\}$, we have*

$$var(H_N(\theta, E_N, X_N)) \leq \frac{C}{N\, s_N^k},$$
$$var(H'_{Ni}(\theta, E_N, X_N)) \leq \frac{C}{N\, s_N^{k+2}},$$
$$var(H''_{Nij}(\theta, E_N, X_N)) \leq \frac{C}{N\, s_N^{k+4}}.$$

These lemmata provide at once pointwise convergence of $H_N(\theta, E_N)$ to $h(\theta)$ in probability for all θ. But pointwise convergence is not enough. We need uniform convergence at least on compact subsets of $\mathbb{R}^q$.

Lemma 6 *Under the conditions [1] - [7] for all compact sets $\Theta \subset \mathbb{R}^q$ and all $\epsilon > 0$, $\delta > 0$, there exists an integer $N_{\Theta,\epsilon,\delta} \in \mathbb{N}$ with*

$$a)\ P\left(\sup_{\theta\in\Theta} |H_N(\theta, E_N, X_N) - h(\theta)| > \delta\right) < \epsilon,$$
$$b)\ P\left(\sup_{\theta\in\Theta} |H'_{Ni}(\theta, E_N, X_N) - h'_i(\theta)| > \delta\right) < \epsilon,$$
$$c)\ P\left(\sup_{\theta\in\Theta} |H''_{Nij}(\theta, E_N, X_N) - h''_{ij}(\theta)| > \delta\right) < \epsilon,$$
$$d)\ P\left(\sup_{\theta\in\Theta} |\lambda_{\max}(H''_N(\theta, E_N, X_N)) - \lambda_{\max}(h''(\theta))| > \delta\right) < \epsilon$$

for all $N \geq N_{\Theta,\epsilon,\delta}$ and $i, j = 1, \ldots, q$.

Since we have uniform convergence according to Lemma 6 the heights of the local maxima of $H_N(\cdot, E_N, X_N)$ converges to the heights of the local maxima of h so that the largest local maxima of $H_N(\cdot, E_N, X_N)$ converges to the largest local maxima of h and thus to the local maxima related to real clusters.

6 Application on edge identification in noisy images

As an application of the cluster method, we use it to detect edges in noisy images. We first use a generalized version of the Rotational Density Kernel Estimator (RDKE) introduced

by Qiu (1997) to estimate those pixels, which may belong to one of the edges, which correspond to the regression lines (hyperplanes) in our model. Then, these points are used as observations z_{nN}.

We choose the RDKE-method because it does not only estimate the points lying on the edges like other methods do, but also the direction of the jump curve in these points. This provides canonical start values for the Newton Raphson method, namely the lines given by the estimated points and directions, which we used instead of those given by any two observations (see the remark after (2) in Section 4).

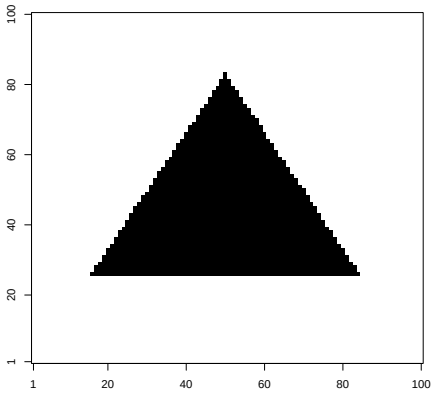

Figure 3: Original image with 100×100 $\{0, 1\}$ - pixels

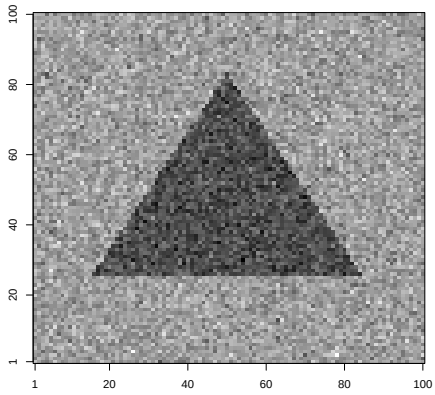

Figure 4: Image with normal distributed noise ($\sigma = 0.3$)

Figure 3 and 4 show the original image with 100×100 pixels with values 0 or 1 and the noisy image with $N(0, 0.3^2)$ - distributed errors.

For using the RDKE - method, we have to choose the window size h_N. In cases like this, where the size of the object which should be detected is known, the window size can be choosen relatively to this size, e.g. 10% of it. The triangle in our example has a diameter of 68 pixels, so we set $h_N = 6.8$.

For every pixel of the image, we used a multiple test ($\alpha = 0.05$) by using the RDKE-method with a uniform kernel for four different angles. Figure 5 shows the 2199 points, for which the hypothesis that the point does not lie on a jump curve could be rejected. We see that there is a large dispersion around the true edges. To avoid boundary problems

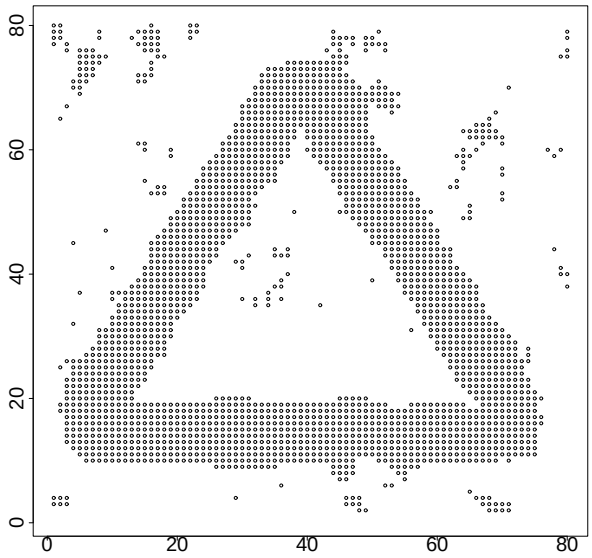

Figure 5: Estimated jump points, respectively observations $z_{n,2199}$

we only used the points with a sufficiently large distance to the border.

On these points, we applicate our estimator with the density of the standard normal distribution as score function ρ. Again, we have to choose a bandwidth – the scale parameter s_N – what could be done with respect to the window size h_N of the RDKE-method. The smaller s_N the more different lines are found. But since it is relatively easy to reduce the number of lines afterwards, better results are achieved if s_N is not too big. Otherwise we would get less lines but each with a bigger deviation. Therefore we choose the scale parameter that way that those points, which lie in the corresponding window of the RDKE-method, have $95\% = 1 - \alpha$ of the weight, that is $s_N = \frac{h_N}{u_{1-\alpha/2}} = \frac{6.8}{u_{1-\alpha/2}} \approx 3.47$. Since in general the window size h_N is not known we also used scale parameters $s_N = 1.5$ and $s_N = 5$ to show the dependence of the choice of the scale parameter. Figure 6 shows the estimated center lines for these three scale parameters.

The number of true clusters is unknown in many applications. Our method provides an easy way to obtain this number, since in general the maxima of the true clusters are considerably larger than the others (see Figure 8). This is contrary to the results of Arslan (2002) where simultaneously the scale parameter is estimated. In this approach the global maximum is attained at a regression line approximating all data as good as possible and

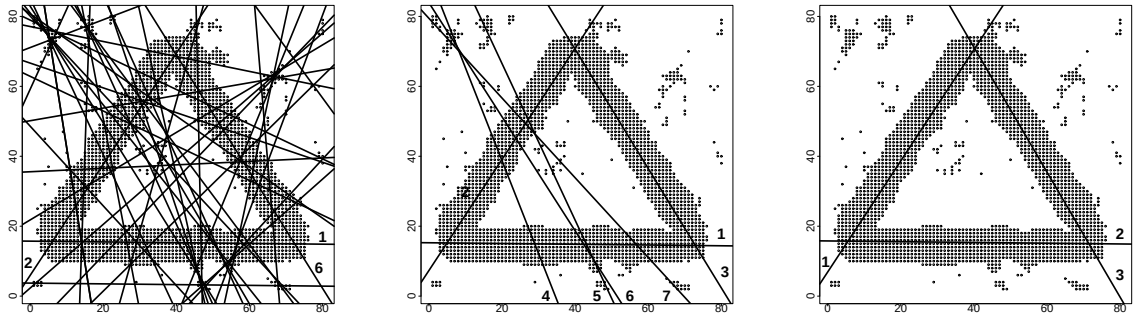

Figure 6: Observations z_{2199} with the estimated cluster lines $M_{2199}(z_{2199})$ for $s_N = 1.5$ (left), $s_N = 3.47$ (middle) and $s_N = 5$ (right)

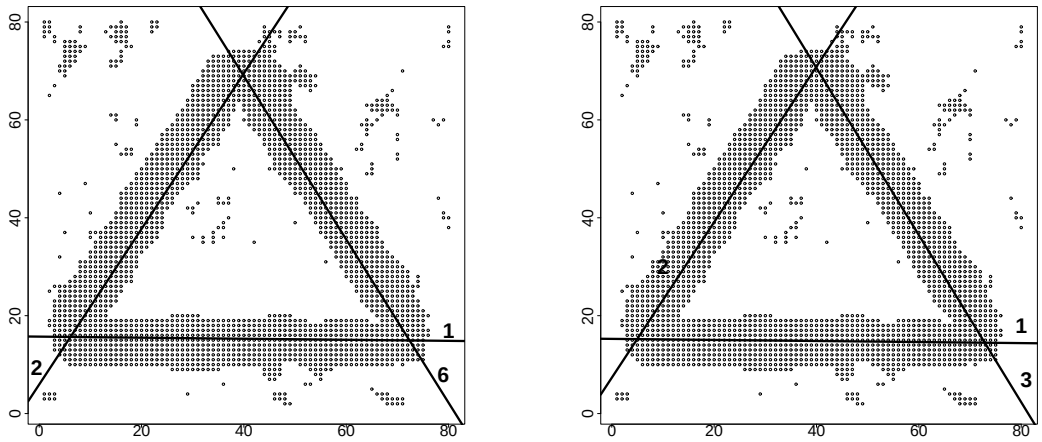

Figure 7: Observations with the three center lines with the largest maxima for $s_N = 1.5$ (left) and $s_N = 3.47$ (right)

not at lines of clusters. This shows that the simultaneous estimation of regression and scale parameter provides problems in choosing the right lines, a reason why Arslan deals with special tests for identifying the true clusters. The problems are caused by the fact that the simultaneous scale estimation provides too large scale estimates which are not converging to zero with growing sample size.

If the number of clusters is known there are several methods to choose the true clusters out of the set $M_N(z_N)$. Beside the usual methods like choosing those clusters to which most of the points belong to, our method suggests also in this case to choose the clusters with the largest local maxima of $H_N(a, b, z_N)$. In this simulation the three clusters with the most points and those with the largest maximum are the same for all three scale parameteres we used (see Figures 7,8,9). Therefore, the choice of the scale parameter

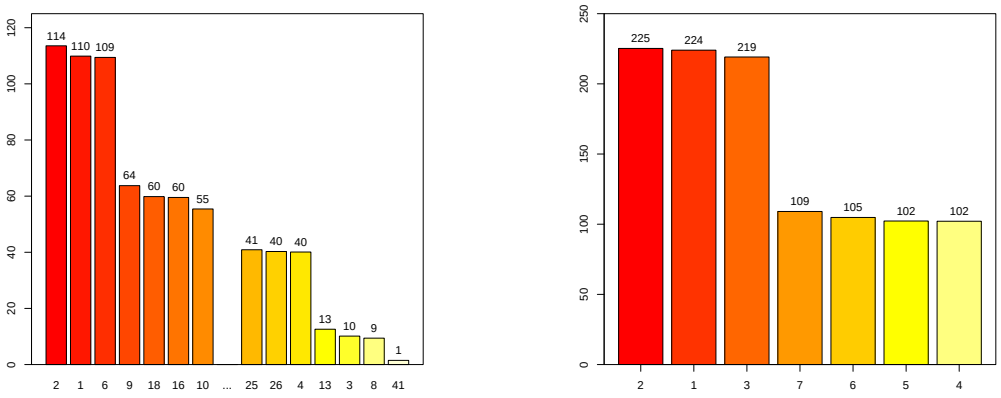

Figure 8: The maxima of the estimated clusters for $s_N = 1.5$ (left) and $s_N = 3.47$ (right)

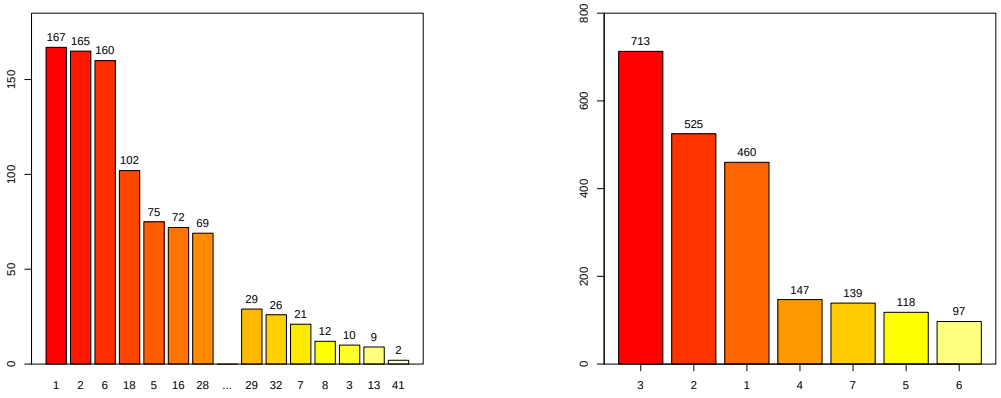

Figure 9: Number of points in the estimated clusters for $s_N = 1.5$ (left) and $s_N = 3.47$ (right)

seems not to be very critical especially if the number of clusters is known.

The estimator provided in this paper has advantages over other methods especially if the clusters are not at all seperated like in this example. For example the regression fixed point cluster (FPC) method introduced by Hennig (2002, 2003) cannot find all three clusters if the "tuning constant", which describes the separation of the clusters, is choosen automatically (see Figure 10). If the method is manually tuned then it finds a fourth cluster containing all points (see Figure 11) but without providing a canonical method to choose the right clusters. Beside that, this method is not independent with respect to rotation (see Figure 12).

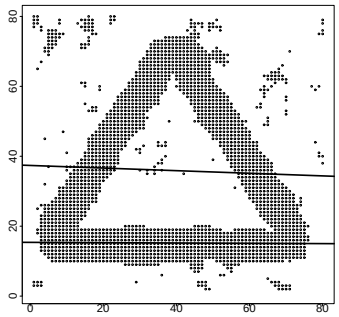

Figure 10: Two cluster center lines found by the FPC method with automatically choosen tuning constant $c = 5.54$

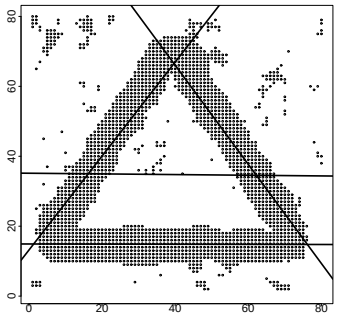

Figure 11: Four cluster center lines found by the FPC method with manually choosen tuning constant $c = 4.0$

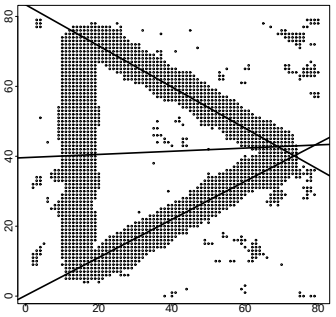

Figure 12: Three cluster center lines found by the FPC method with manually choosen tuning constant $c = 4.0$ and rotated observations

7 Appendix: Proofs

For simplicity we dropped in the lemmata some assumptions. From the proofs it is clear which of the assumptions in Section 5 are needed for the particular lemmata.

Proof of Lemma 1. Let $\beta \neq \beta_0$. Then we obtain for the derivative of h

$$\begin{aligned}
\frac{\partial}{\partial\beta} h(\beta)\,(\beta-\beta_0) &= \int f'((1,x^\top)(\beta-\beta_0))\,(1,x^\top)\,G_o(dx)\,(\beta-\beta_0) \\
&= \int_{(1,x^\top)(\beta-\beta_0)>0} f'((1,x^\top)(\beta-\beta_0))\,(1,x^\top)(\beta-\beta_0)\,G_o(dx) \\
&\quad + \int_{(1,x^\top)(\beta-\beta_0)<0} f'((1,x^\top)(\beta-\beta_0))\,(1,x^\top)(\beta-\beta_0)\,G_o(dx) \\
&< 0.
\end{aligned}$$

Hence h has only a maximum at $\beta = \beta_0$. □

Proof of Lemma 2. Assume that G_j has positive mass on a subspace S of $\mathbb{R}^{p-1}$. Then there exists $\tilde{\beta}$ with $\left(1, x^\top\right)\tilde{\beta} = 0$ for all $x \in S$. This means

$$h(v\tilde{\beta}) = \sum_{l=1}^{L} \gamma_l \int f((1,x^\top)(v\tilde{\beta}-\beta_l))\,G_l(dx)$$

$$\geq \quad \gamma_j \int_S f(-(1,x^\top)\beta_j)\, G_j(dx) =: w$$

for all $v \in I\!R$. If $\eta \in I\!N$ is such that $w \geq \frac{1}{\eta}$ then $\{v\tilde{\beta};\ v \in I\!R\}$ is a subset of Θ_η so that Θ_η is not compact. Conversely, assume that none of distributions G_l has positive mass on a subspace of $I\!R^{p-1}$. Let $\eta \in I\!N$ arbitrary and regard any sequence $\beta^{(n)} \in \Theta_\eta$. First assume that $\beta^{(n)}$ is unbounded. The only possibility to obtain $\limsup_{n\to\infty} f\left(\left(1,x^\top\right)\left(\beta^{(n)}-\beta_l\right)\right) > 0$ is that $\left(1,x^\top\right)\beta^{(n)}$ is bounded and this can be only satisfied by x from a subspace of $I\!R^{p-1}$. Since G_l has no positive mass on a subspace we have $\lim_{n\to\infty} f\left(\left(1,x^\top\right)\left(\beta^{(n)}-\beta_l\right)\right) = 0$ G_l-almost surely for all $l = 1,\ldots,L$. This implies

$$\lim_{n\to\infty} h(\beta^{(n)}) = \lim_{n\to\infty} \sum_{l=1}^{L} \gamma_l \int f\left(\left(1,x^\top\right)\left(\beta^{(n)}-\beta_l\right)\right)\, G_l(dx) = 0$$

and therefore a contradiction to $\beta^{(n)} \in \Theta_\eta$. Hence $\beta^{(n)}$ is bounded. A subsequence of $\beta^{(n)}$ converges also to a member $\beta^{(0)}$ of Θ_η since, with f, also h is continuous. Hence Θ_η is compact. □

Proof of Lemma 3. Let F be the distribution function belonging to the density f. Since $a^\top(E_{1nN}^\top, E_{2nN})^\top$ has the density f and $Y_{nN} = \frac{b_0}{a_{0p}} - \frac{1}{a_{0p}} a_{01}^\top X_{nN}$ we have

$$\begin{aligned}
&P(a_1^\top V_{nN} + a_p W_{nN} \leq b) \\
&\quad = \quad P\left(a_1^\top X_{nN} + a_p Y_{nN} + a^\top (E_{1nN}^\top, E_{2nN})^\top \leq b\right) \\
&\quad = \quad P\left(a_1^\top X_{nN} + a_p \frac{b_0}{a_{0p}} - \frac{a_p}{a_{0p}} a_{01}^\top X_{nN} + a^\top (E_{1nN}^\top, E_{2nN})^\top \leq b\right) \\
&\quad = \quad \int P\left(a_p \frac{b_0}{a_{0p}} + \left(a_1^\top - \frac{a_p}{a_{0p}} a_{01}^\top\right) x + a^\top (E_{1nN}^\top, E_{2nN})^\top \leq b \;\middle|\; X_{nN} = x\right) G_0(dx) \\
&\quad = \quad \int P\left(a^\top (E_{1nN}^\top, E_{2nN})^\top \;\leq\; b - \frac{a_p}{a_{0p}} b_0 - \left(a_1^\top - \frac{a_p}{a_{0p}} a_{01}^\top\right) x \;\middle|\; X_{nN} = x\right) G_0(dx) \\
&\quad = \quad \int F\left(b - \frac{a_p}{a_{0p}} b_0 - \left(a_1^\top - \frac{a_p}{a_{0p}} a_{01}^\top\right) x\right)\; G_0(dx).
\end{aligned}$$

Differentiation with respect to b provides the assertion. □

Proof of Lemma 4. We prove the assertion for the most difficult case, that is for $E(H''_{Nij}(\theta, E_N, X_N))$. The other cases follow similarly.

Since $\rho(e) \to 0$ for $e_r \to \pm\infty$ we have in particular $\rho\left(\frac{e-c^l(\theta,x)}{s_N}\right) f'(e) \to 0$ and

$\rho_s' \left(\frac{e-c^l(\theta,x)}{s_N} \right) f(e) \to 0$ for $e_r \to \pm\infty$, $r,s = 1,\ldots,k$. Hence partial integration provides

$$
\begin{aligned}
E(H''_{Nij}(\theta, E_N, X_N)) &= \sum_{l=1}^{L} \gamma_l \int \frac{1}{s_N^k} \frac{\partial^2}{\partial\theta_i \partial\theta_j} \rho \left(\frac{e - c^l(\theta,x)}{s_N} \right) f(e)\, de\, G_l(dx) \\
&= \sum_{l=1}^{L} \gamma_l \int \frac{1}{s_N^k} \frac{\partial}{\partial\theta_i} \left(\rho' \left(\frac{e - c^l(\theta,x)}{s_N} \right)^\top \frac{\partial}{\partial\theta_j} \frac{-c^l(\theta,x)}{s_N} \right) f(e)\, de\, G_l(dx) \\
&= \sum_{l=1}^{L} \gamma_l \int \frac{1}{s_N^k} \left[\frac{\partial}{\partial\theta_i} \frac{c^l(\theta,x)^\top}{s_N} \rho'' \left(\frac{e - c^l(\theta,x)}{s_N} \right) \frac{\partial}{\partial\theta_j} \frac{c^l(\theta,x)}{s_N} \right. \\
&\quad \left. - \rho' \left(\frac{e - c^l(\theta,x)}{s_N} \right)^\top \frac{\partial^2}{\partial\theta_i\, \partial\theta_j} \frac{c^l(\theta,x)}{s_N} \right] f(e)\, de\, G_l(dx) \\
&= \sum_{l=1}^{L} \gamma_l \int \frac{1}{s_N^k} \left[\frac{\partial}{\partial\theta_i} c^l(\theta,x)^\top \frac{\partial^2}{\partial^2 e} \rho \left(\frac{e - c^l(\theta,x)}{s_N} \right) \frac{\partial}{\partial\theta_j} c^l(\theta,x) \right. \\
&\quad \left. - \frac{\partial}{\partial e} \rho \left(\frac{e - c^l(\theta,x)}{s_N} \right)^\top \frac{\partial^2}{\partial\theta_i\, \partial\theta_j} c^l(\theta,x) \right] f(e)\, de\, G_l(dx) \\
&= \sum_{l=1}^{L} \gamma_l \int \frac{1}{s_N^k} \left[\int \rho \left(\frac{e - c^l(\theta,x)}{s_N} \right) \frac{\partial}{\partial\theta_i} c^l(\theta,x)^\top \frac{\partial^2}{\partial^2 e} f(e) \frac{\partial}{\partial\theta_j} c^l(\theta,x)\, de \right. \\
&\quad \left. + \int \rho \left(\frac{e - c^l(\theta,x)}{s_N} \right) \left(\frac{\partial}{\partial e} f(e) \right)^\top \frac{\partial^2}{\partial\theta_i\, \partial\theta_j} c^l(\theta,x)\, de \right] G_l(dx).
\end{aligned}
$$

Substituting $v = \frac{e - c^l(\theta,x)}{s_N}$ we get

$$
\begin{aligned}
&E(H''_{Nij}(\theta, E_N, X_N)) \\
&\quad = \sum_{l=1}^{L} \gamma_l \int \left[\int \rho(v) \frac{\partial}{\partial\theta_i} c^l(\theta,x)^\top f'' \left(s_N\, v + c^l(\theta,x) \right) \frac{\partial}{\partial\theta_j} c^l(\theta,x)\, dv \right. \\
&\qquad \left. + \int \rho(v) f' \left(s_N\, v + c^l(\theta,x) \right)^\top \frac{\partial^2}{\partial\theta_i\, \partial\theta_j} c^l(\theta,x)\, dv \right] G_l(dx) \\
&\quad = \sum_{l=1}^{L} \gamma_l \int \int \rho(v) \frac{\partial^2}{\partial\theta_i\, \partial\theta_j} f \left(s_N\, v + c^l(\theta,x) \right) dv\, G_l(dx).
\end{aligned}
$$

This implies because of $\int \rho(v) dv = 1$, $\int \rho(v) \|v\| dv < \infty$ and the Lipschitz continuity of f'' and $\frac{\partial^2}{\partial\theta_i\theta_j} c^l(\theta,x)$

$$
\begin{aligned}
&|E(H''_{Nij}(\theta, E_N, X_N)) - h''_{ij}(\theta)| \\
&\quad \le \sum_{l=1}^{L} \gamma_l \int \int \rho(v) \left| \frac{\partial^2}{\partial\theta_i\, \partial\theta_j} f \left(s_N\, v + c^l(\theta,x) \right) - \frac{\partial^2}{\partial\theta_i\, \partial\theta_j} f \left(c^l(\theta,x) \right) \right| dv\, G_l(dx) \\
&\quad \le \sum_{l=1}^{L} \gamma_l \int \int \rho(v)\, C_0\, s_N\, |v|\, dv\, G_l(dx) = C s_N,
\end{aligned}
$$

where the constants C_0 and C are independent of θ and N. □

Proof of Lemma 5. Again, we prove the assertion for the most difficult case, that is for $\mathrm{var}(H''_{Nij}(\theta, E_N, X_N))$. The other cases follow similarly. With Lemma 4, substitution of $v = \frac{e - c^l(\theta,x)}{s_N}$, and $\sup_y f(y) < \infty$ because of Condition [3], we obtain

$$
\begin{aligned}
&\mathrm{var}(H''_{Nij}(\theta, E_N, X_N)) \\
&= \frac{1}{N}\left[\sum_{l=1}^{L}\gamma_l \int \frac{1}{s_N^{2k}}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\,\rho\left(\frac{e-c^l(\theta,x)}{s_N}\right)\right)^2 f(e)\,de\,G_l(dx)\right. \\
&\quad \left. - \left(E(H''_{Nij}(\theta, E_N, X_N))\right)^2\right] \\
&= \frac{1}{N}\sum_{l=1}^{L}\gamma_l \int \frac{1}{s_N^{2k}}\left(\frac{\partial}{\partial\theta_i}\left(\rho'\left(\frac{e-c^l(\theta,x)}{s_N}\right)^\top \frac{\partial}{\partial\theta_j}\frac{-c^l(\theta,x)}{s_N}\right)\right)^2 f(e)\,de\,G_l(dx) \\
&\quad -\frac{1}{N}\left(h''_{ij}(\theta) + O(s_N)\right)^2 \\
&= \frac{1}{N}\sum_{l=1}^{L}\gamma_l \int \frac{1}{s_N^{2k}}\left[\frac{\partial}{\partial\theta_i}\frac{c^l(\theta,x)^\top}{s_N}\,\rho''\left(\frac{e-c^l(\theta,x)}{s_N}\right)\frac{\partial}{\partial\theta_j}\frac{c^l(\theta,x)}{s_N}\right. \\
&\quad \left. -\rho'\left(\frac{e-c^l(\theta,x)}{s_N}\right)^\top \frac{\partial^2}{\partial\theta_i\partial\theta_j}\frac{c^l(\theta,x)}{s_N}\right]^2 f(e)\,de\,G_l(dx) \;+\; O\left(\frac{1}{N}\right) \\
&= \frac{1}{N}\sum_{l=1}^{L}\gamma_l \int \frac{1}{s_N^{k}}\left[\frac{\partial}{\partial\theta_i}\frac{c^l(\theta,x)^\top}{s_N}\,\rho''(v)\,\frac{\partial}{\partial\theta_j}\frac{c^l(\theta,x)}{s_N}\right. \\
&\quad \left. -\rho'(v)^\top \frac{\partial^2}{\partial\theta_i\partial\theta_j}\frac{c^l(\theta,x)}{s_N}\right]^2 f\left(s_N\,v + c^l(\theta,x)\right)\,dv\,G_l(dx) \;+\; O\left(\frac{1}{N}\right) \\
&\le \frac{1}{N\,s_N^{k+4}}\sum_{l=1}^{L}\gamma_l \int \left[\frac{\partial}{\partial\theta_i}c^l(\theta,x)^\top\,\rho''(v)\,\frac{\partial}{\partial\theta_j}c^l(\theta,x)\right. \\
&\quad \left. -s_N\,\rho'(v)^\top \frac{\partial^2}{\partial\theta_i\partial\theta_j}c^l(\theta,x)\right]^2 C_1\,dv\,G_l(dx) \;+\; O\left(\frac{1}{N}\right)
\end{aligned}
$$

with constant C_1 independent of θ. The conditions $\int \rho''_{rs}(v)^2\,dv < \infty$, $\int \rho'_r(v)^2\,dv < \infty$, $\sup_{\theta\in\Theta, x\in\mathcal{X}}\left|\frac{\partial}{\partial\theta_i}c^l_r(\theta,x)\right| < \infty$, $\sup_{\theta\in\Theta, x\in\mathcal{X}}\left|\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}c^l_r(\theta,x)\right| < \infty$ for $r, s = 1, \ldots, k$, for $r, s = 1, \ldots, k$ and $l = 1, \ldots, L$, provide

$$
\mathrm{var}(H''_{Nij}(\theta, E_N, X_N)) \le \frac{C}{N\,s_N^{k+4}}
$$

where the constant C is independent of θ and N. □

Proof of Lemma 6. The representation of the largest eigenvalue of a matrix $A \in I\!R^{q\times q}$ as $\lambda_{\max}(A) = \sup_{\|a\|=1} a^\top A\, a$ provides that the largest eigenvalue is a uniform continuous mapping from $I\!R^{q\times q}$ to $I\!R$. Hence the assertion d) follows from c). Since the proofs for a) and b) are similar to c), we prove again the assertion only for H''_{Nij}. Fix an arbitrary $\delta > 0$, $\epsilon > 0$, and a compact set $\Theta \subset I\!R^q$. Since f is a density and ρ can be interpreted as a density, the Lipschitz continuity of the derivatives of f and ρ implies that the derivatives are bounded. Choose $c > 1$ such that c is Lipschitz constant for f'', ρ'', $\frac{\partial}{\partial\theta_i} c_r^l(\cdot, x)$, $\frac{\partial^2}{\partial\theta_i\,\partial\theta_j} c_r^l(\cdot, x)$ and such that $\sup_{y\in I\!R^k} ||f'(y)|| < c$, $\sup_{y\in I\!R^k} ||f''(y)|| < c$, $\sup_{y\in I\!R^k} ||\rho'(y)|| < c$, $\sup_{y\in I\!R^k} ||\rho''(y)|| < c$, $\sup_{\theta\in\Theta, x\in\mathcal{X}} \left|\frac{\partial}{\partial\theta_i} c_r^l(\theta, x)\right| < c$, and $\sup_{\theta\in\Theta, x\in\mathcal{X}} \left|\frac{\partial^2}{\partial\theta_i\,\partial\theta_j} c_r^l(\theta, x)\right| < c$ for $r, s = 1, \ldots, k$, $i, j = 1, \ldots, q$, $l = 1, \ldots, L$. Since $\Theta \subset I\!R^q$ is compact, there exists $D \in I\!R$ and $M_N := \left\lceil \frac{D}{s_N^{q(k+3)}} \right\rceil$ disjoint subsets $J_{1N}, \ldots, J_{M_N N}$ of Θ such that

$$\Theta = \bigcup_{m=1}^{M_N} J_{mN},$$

$$\sup_{\theta,\theta'\in J_{mN}} \|\theta - \theta'\| \le \frac{\delta}{18\, c^4}\, s_N^{k+3}$$

for all $m = 1, \ldots, M_N$, $N \in I\!N$. Let $\theta_{mN} \in J_{mN}$ for $m = 1, \ldots, M_N$. According to Lemma 4 and Lemma 5, there exists c_1 and $N_0 \in I\!N$ such that

$$\begin{aligned}
&P\left(|H''_{Nij}(\theta_{mN}, E_N, X_N) - h''_{ij}(\theta_{mN})| > \frac{\delta}{3}\right) \\
&\le\ P\left(|H''_{Nij}(\theta_{mN}, E_N, X_N) - E(H''_{Nij}(\theta_{mN}, E_N, X_N))| > \frac{\delta}{6}\right. \\
&\qquad\qquad \left. \text{or } |E(H''_{Nij}(\theta_{mN}, E_N, X_N)) - h''_{ij}(\theta_{mN})| > \frac{\delta}{6}\right) \\
&\le\ \frac{36}{\delta^2} \operatorname{var}\left(H''_{Nij}(\theta_{mN}, E_N, X_N)\right) \\
&\le\ \frac{36}{\delta^2}\, c_1\, \frac{1}{N\, s_N^{k+4}} \\
&\le\ \frac{36\, c_1\, 2\, D}{\delta^2\, \epsilon\, N\, s_N^{q(k+3)+k+4}}\, \frac{\epsilon}{2\, D}\, s_N^{q(k+3)} \\
&\le\ \frac{\epsilon}{2\, D}\, s_N^{q(k+3)}
\end{aligned}$$

for all $N \ge N_0$ and $m = 1, \ldots, M_N$, since $s_N \to 0$ and $N\, s_N^{q(k+3)+k+4} \to \infty$. Then for $N \ge N_0$, we have

$$\sup_{m=1,\ldots,M_N} \left(P(|H''_{Nij}(\theta_{mN}, E_N, X_N) - h''_{ij}(\theta_{mN})| > \frac{\delta}{3}\right)$$

$$
\begin{aligned}
&\leq \sum_{m=1}^{M_N} \left(P(|H''_{Nij}(\theta_{mN}, E_N, X_N) - h''_{ij}(\theta_{mN})| > \frac{\delta}{3} \right) \\
&\leq M_N \, \frac{\epsilon}{2\,D} \, s_N^{q(k+3)} \\
&\leq \left\lceil \frac{D}{s_N^{q(k+3)}} \right\rceil \, \frac{\epsilon}{2\,D} \, s_N^{q(k+3)} < \epsilon.
\end{aligned}
$$

The Lipschitz continuity and bounds of f', f'', ρ', ρ'', $\frac{\partial}{\partial\theta_i} c_r^l$, $\frac{\partial^2}{\partial\theta_i\,\partial\theta_j} c_r^l$ provide for all $\theta \in J_{mN}$, $m = 1, \ldots, M_N$, $e_N = (e_{1N}, \ldots, e_{NN})^\top \in I\!R^{N\times k}$, $x_N = (x_{1N}, \ldots, x_{NN})^\top \in \mathcal{X}^N$

$$
\begin{aligned}
&|H''_{Nij}(\theta_{mN}, e_N, x_N) - H''_{Nij}(\theta, e_N, x_N)| \\
&\leq \frac{1}{N} \sum_{n=1}^{N} \frac{1}{s_N^{k+2}} \left| \frac{\partial}{\partial\theta_i} c^n(\theta_{mN}, x_{nN})^\top \, \rho'' \left(\frac{e_{nN} - c^n(\theta_{mN}, x_{nN})}{s_N} \right) \, \frac{\partial}{\partial\theta_j} c^n(\theta_{mN}, x_{nN}) \right. \\
&\quad \left. - \frac{\partial}{\partial\theta_i} c^n(\theta, x_{nN})^\top \, \rho'' \left(\frac{e_{nN} - c^n(\theta, x_{nN})}{s_N} \right) \, \frac{\partial}{\partial\theta_j} c^n(\theta, x_{nN}) \right| \\
&\quad + \frac{1}{N} \sum_{n=1}^{N} \frac{1}{s_N^{k+1}} \left| \rho' \left(\frac{e_{nN} - c^n(\theta_{mN}, x_{nN})}{s_N} \right)^\top \, \frac{\partial^2}{\partial\theta_i \partial\theta_j} c^n(\theta_{mN}, x_{nN}) \right. \\
&\quad \left. - \rho' \left(\frac{e_{nN} - c^n(\theta, x_{nN})}{s_N} \right)^\top \, \frac{\partial^2}{\partial\theta_i \partial\theta_j} c^n(\theta, x_{nN}) \right| \\
&\leq \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{s_N^{k+2}} \left(2c^3 + \frac{1}{s_N} c^4 \right) + \frac{1}{s_N^{k+1}} \left(c^2 + \frac{1}{s_N} c^3 \right) \right) \, \|\theta_{mN} - \theta\| \\
&\leq \frac{1}{N} \sum_{n=1}^{N} \frac{1}{s_N^{k+3}} \, 6\, c^4 \, \|\theta_{mN} - \theta\| < \frac{\delta}{3},
\end{aligned}
$$

and similarly

$$
|h''_{ij}(\theta_{mN}) - h''_{ij}(\theta)| < \frac{\delta}{3}.
$$

Therefore, for all (e_N, x_N) with

$$
\sup_{m=1,\ldots,M_N} |H''_{Nij}(\theta_{mN}, e_N, x_N) - h''_{ij}(\theta_{mN})| \leq \frac{\delta}{3},
$$

we have

$$
\begin{aligned}
&\sup_{\theta\in\Theta} |H''_{Nij}(\theta, e_N, x_N) - h''_{ij}(\theta)| \\
&\quad = \sup_{m=1,\ldots,M_N} \sup_{\theta\in J_{mN}} |H''_{Nij}(\theta, e_N, x_N) - h''_{ij}(\theta)| \\
&\quad = \sup_{m=1,\ldots,M_N} \sup_{\theta\in J_{mN}} \big| H''_{Nij}(\theta, e_N, x_N) - H''_{Nij}(\theta_{mN}, e_N, x_N) \\
&\qquad + H''_{Nij}(\theta_{mN}, e_N, x_N) - h''_{ij}(\theta_{mN}) + h''_{ij}(\theta_{mN}) - h''_{ij}(\theta) \big| \\
&\quad \leq \delta.
\end{aligned}
$$

This implies

$$\begin{aligned} & P\left(\sup_{\theta\in\Theta}|H''_{Nij}(\theta,E_N,X_N)-h''_{ij}(\theta)|\le\delta\right) \\ \ge\ & P\left(\sup_{m=1,\dots,M_N}|H''_{Nij}(\theta_{mN},E_N,X_N)-h''_{ij}(\theta_{mN})|\le\frac{\delta}{3}\right) \\ \ge\ & 1-\epsilon, \end{aligned}$$

and thus the assertion. □

Proof of the Theorem 1. Let $\delta,\epsilon>0$ arbitrary. At first we show $P(\mathcal{M}_N^\eta(E_N,X_N)\subset\mathcal{U}_\delta(\mathcal{M}))\ge 1-\frac{\epsilon}{2}$ for all $\eta\in I\!N$, and then $P(\mathcal{M}\subset\mathcal{U}_\delta(\mathcal{M}_N^\eta(E_N,X_N)))\ge 1-\frac{\epsilon}{2}$ for $\eta\ge\eta_0$.

1. Condition [10] ensures that there exists $\epsilon_1>0$ such that $|\lambda_{\max}(h''(\theta))|>\epsilon_1$ for all $\theta\in\mathcal{M}_0\cap\Theta_\eta$, where Θ_η is defined by (3) and $\mathcal{M}_0$ by [10]. Since, with f'' and $\frac{\partial^2}{\partial\theta_i\partial\theta_j}c^l(\theta,x)$, h'' is continuous and thus uniformly continuous on the compact set Θ_η, there exists $\epsilon_2>0$, $\delta_1>0$ such that $\delta_1\le\delta$ and $|\lambda_{\max}(h''(\theta))|>\epsilon_2$ for all $\theta\in\mathcal{U}_{\delta_1}(\mathcal{M}_0)\cap\Theta_\eta$.

Since $h'(\theta)\neq 0$ for all $\theta\in\Theta_\eta\setminus\mathcal{M}_0$ and $\Theta_\eta\setminus\mathcal{U}_{\delta_1}(\mathcal{M}_0)$ is compact, the continuity of h' implies the existence of $\epsilon_3>0$ with $\|h'(\theta)\|>\epsilon_3$ for all $\theta\in\Theta_\eta\setminus\mathcal{U}_{\delta_1}(\mathcal{M}_0)$. This can be shown by contradiction: Assume for all $\kappa\in I\!N$ there exists $\theta_\kappa\in\Theta_\eta\setminus\mathcal{U}_{\delta_1}(\mathcal{M}_0)$ with $\|h'(\theta_\kappa)\|\le\frac{1}{\kappa}$. The compactness of $\Theta_\eta\setminus\mathcal{U}_{\delta_1}(\mathcal{M}_0)$ implies $\lim_{\kappa\to\infty}\theta_\kappa=\theta_0\in\Theta_\eta\setminus\mathcal{U}_{\delta_1}(\mathcal{M}_0)$. However, since h and h' are continuous, we have $\frac{1}{\eta}\le\lim_{\kappa\to\infty}h(\theta_\kappa)=h(\theta_0)$ and $0=\lim_{\kappa\to\infty}h'(\theta_\kappa)=h'(\theta_0)$. This means that $\theta_0\in\mathcal{U}_{\delta_1}(\mathcal{M}_0)$ which is a contradiction.

Now let A_N denote the set of all $e_N\in I\!R^{N\times k}$ and $x_N\in\mathcal{X}^N$ with $\sup_{\theta\in\Theta_\eta}\|H'_N(\theta,e_N,x_N)-h'(\theta)\|<\epsilon_3/2$ and $\sup_{\theta\in\Theta_\eta}|\lambda_{\max}(H''_N(\theta,e_N,x_N))-\lambda_{\max}(h''(\theta))|<\epsilon_2/2$. Then for all $(e_N,x_N)\in A_N$, we have:

$$\begin{aligned} &\|H'_N(\theta,e_N,x_N)\|>\epsilon_3/2 \text{ for all } \theta\in\Theta_\eta\setminus\mathcal{U}_{\delta_1}(\mathcal{M}_0),\\ &\lambda_{\max}(H''_N(\theta,e_N,x_N))<-\epsilon_2/2 \text{ for all } \theta\in\mathcal{U}_{\delta_1}(\mathcal{M}_0)\cap\Theta_\eta \text{ with } \lambda_{\max}(h''(\theta))<0,\\ &\lambda_{\max}(H''_N(\theta,e_N,x_N))>\epsilon_2/2 \text{ for all } \theta\in\mathcal{U}_{\delta_1}(\mathcal{M}_0)\cap\Theta_\eta \text{ with } \lambda_{\max}(h''(\theta))>0. \end{aligned}$$

Hence, the local extrema of $H_N(\cdot,e_N,x_N)$ within Θ_η are all lying in $\mathcal{U}_{\delta_1}(\mathcal{M}_0)$. Moreover, the local maxima are lying in sets $\mathcal{U}_{\delta_1}(\{\theta\})$ where θ is a maximum point of h. This implies $\mathcal{M}_N(e_N,x_N)\cap\Theta_\eta\subset\mathcal{U}_{\delta_1}(\mathcal{M})\subset\mathcal{U}_\delta(\mathcal{M})$ and thus

$$A_N\subset\{(e_N,x_N);\ \mathcal{M}_N(e_N,x_N)\cap\Theta_\eta\subset\mathcal{U}_\delta(\mathcal{M})\}.$$

Since Lemma 6 implies the existence of N_0 such that $P(A_N) \geq 1 - \frac{\epsilon}{2}$ for all $N \geq N_0$ the first assertion is proved.

2. Since $\mathcal{M}$ is finite because of [9], there exists η_0 such that $\mathcal{M}$ is included in an open subset of Θ_{η_0}. Moreover, Condition [9] provides the existence of $0 < \delta_1 < \delta_2 < \delta$ and $\epsilon_1 > 0$ with $h(\theta) < h(\theta_0) - \epsilon_1$ for all $\theta \in \mathcal{U}_{\delta_2}(\mathcal{M}) \setminus \mathcal{U}_{\delta_1}(\mathcal{M})$ and all $\theta_0 \in \mathcal{M}$. Additionally, δ_2 can be chosen such that $\mathcal{U}_{\delta_2}(\mathcal{M}) \subset \Theta_{\eta_0}$.

For any $\eta > \eta_0$, let A_N denote the set of all $e_N \in I\!R^{N \times k}$ and $x_N \in \mathcal{X}^N$ with $\sup_{\theta \in \Theta_\eta} \|H_N(\theta, e_N, x_N) - h(\theta)\| < \epsilon_1/3$. Then we have for all $(e_N, x_N) \in A_N$

$$H_N(\theta, e_N, x_N) < h(\theta) + \frac{\epsilon_1}{3} < h(\theta_0) - \frac{2\epsilon_1}{3} < H_N(\theta_0, e_N, x_N) - \frac{\epsilon_1}{3}$$

for all $\theta \in \mathcal{U}_{\delta_2}(\mathcal{M}) \setminus \mathcal{U}_{\delta_1}(\mathcal{M})$ and all $\theta_0 \in \mathcal{M}$. This means that $H_N(\theta, e_N, x_N)$ has a local maximum within $\mathcal{U}_{\delta_1}(\{\theta_0\})$ for each $\theta_0 \in \mathcal{M}$. Hence $\mathcal{M} \subset \mathcal{U}_\delta(\mathcal{M}_N^\eta(e_N, x_N))$. Again Lemma 6 provides the existence of N_0 such that $P(A_N) \geq 1 - \frac{\epsilon}{2}$ for all $N \geq N_0$ so that the second assertion is proved. □

Acknowledgements

The authors would like to thank one referee for the very valuable and constructive comments which improved the paper a lot.

References

[1] Arslan, O. (2002). A simple test to identify good solutions of redescending M estimating equations for regression. In: *Developments in Robust Statistics, Proceedings of ICORS 2001*, Dutter, R., Gather, U., Rousseeuw, P.J. and Filzmoser, P. (Eds.), 50-61.

[2] Bednarski, T., Clarke, B.R. and Kolkiewics, W. (1991). Statistical expansions and local uniform Fréchet differentiability. *J. Austral. Math. Soc. Ser. A* **50**, 88-97.

[3] Chen, H. and Meer, P. (2002). Robust computer vision through kernel density estimation. In: *ECCV 2002, LNCS 2350*, A. Heyden et al. (Eds.), Springer, Berlin, 236-250.

[4] Chen, H., Meer, P. and Tyler, D. E. (2001). Robust regression for data with multiple structures. In: *2001 IEEE Conference on Computer Vision and Pattern Recognition*, vol. I, Kauai, HI, 1069-1075.

[5] Chu, C. K., Glad, I. K., Godtliebsen, F., Marron, J. S. (1998). Edge-preserving smoothers for image processing. *J. Amer. Statist. Assoc.* **93**, 526-541.

[6] Davies, P. L. (1988). Consistent estimates for finite mixtures of well separated elliptical distributions. In: *Classification and Related Methods of Data Analysis*, Bock, H.-H. (Ed.), Elsevier Science Publishers, Amsterdam, 195-202.

[7] Desarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification* **5**, 249-282.

[8] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics - The Approach Based on Influence Functions.* John Wiley, New York.

[9] Hennig, C. (1997). *Datenanalyse mit Modellen für Cluster linearer Regression.* Dissertation, Universität Hamburg, Fachbereich Mathematik.

[10] Hennig, C. (2000). Regression fixed point clusters: motivation, consistency and simulations. *Preprint 2000-02,* Universität Hamburg, Fachbereich Mathematik.

[11] Hennig, C. (2003). Clusters, outliers, and regression: Fixed point clusters. *Journal of Multivariate Analysis.* **86/1**, 183-212.

[12] Hillebrand, M. and Müller, Ch.H. (2001). On consistency of redescending M-kernel smoothers. *Submitted.*

[13] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Ann. Statist.* **1**, 799-821.

[14] Huber, P.J. (1981). *Robust Statistics.* John Wiley, New York.

[15] Kiefer, N.M. (1978). Discrete parameter variation: Efficient estimation of a switching regression model. *Econometrica* **46**, 427-434.

[16] Krishnapuram, R. and Freg, C.-P. (1992). Fitting an unknown number of lines and planes to image data through compatible cluster merging. *Pattern Recognition* **25**, 385-400.

[17] Meer, P. and Tyler, D.E. (1998). Smoothing the gap between statistics and image understanding. Comments on the paper *Edge-preserving smoothers for image processing*, Chu, C. K., Glad, I. K., Godtliebsen, F., Marron, J. S., *J. Amer. Statist. Assoc.* **93**, 526-541.

[18] Morgenthaler, S. (1990). Fitting redescending M-estimators in regression. In: *Robust Regression,* Lawrence, H. D. and Arthur, S. (Eds.), Dekker, New York, 105-128.

[19] Müller, Ch.H. (2002). Comparison of high breakdown point estimators for image denoising. *Allg. Stat. Archiv* **86**, 307-321.

[20] Pollard, D. (1981). Strong consistency of K-means clustering. *Ann. Statist.* **9**, 135-140.

[21] Qiu, P. (1997). Nonparametric estimation of jump surface. *The Indian Journal of Statistics* **59**, Series A, 268-294.

[22] Roth, G. and Levine, M.D. (1993). Extracting geometric primitive. *CVGIP: Image Understanding* **58**, 1-22.

[23] Scott, D.W. (1992). *Multivariate density estimation*. Wiley, New York.

[24] Späth, H. (1979). Clusterwise linear regression. *Computing* **22**, 367-373.

[25] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[26] Stewart, C. V. (1997). Bias in robust estimation caused by discontinuities and multiple structures. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **19**, 818-833.

[27] Zamar, R. H. (1989). Robust estimation in the errors-in-variables model. *Biometrika* **76**, 149-160.

mueller@math.uni-oldenburg.de

Fachbereich 6 - Mathematik

Universität Oldenburg

Postfach 2503

D-26111 Oldenburg

Germany