Regression clustering with redescending M-estimators

Tim Garlipp and Christine H. Müller

Universität Oldenburg, Fachbereich 6 Mathematik
Postfach 2503, D - 26111 Oldenburg, Germany

Abstract. We use the local maxima of a redescending M-estimator to identify clusters, 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 locations and give consistency results for all three cases. The approach of orthogonal regression is applied to the identification of edges in noisy images.

1 Introduction

For independently and identically distributed random variables $Y_1, \ldots, Y_N$, the (location-) M-estimator is defined as (global or some local) maximum of

$$H_N(y) = \sum_{n=1}^{N} \rho(Y_n - y).$$

If ρ' is strictly monotone the objective function is unimodal, so that the maximum is unique. For example with $\rho(y) = -y^2$, the maximum is attained at the mean of the observations. But using score functions with redescending derivatives, $H_n(y)$ can have several local maxima, what especially has the disadvantage that computation of the M-estimator is more complicated. But since these local maxima correspond to substructures in the data, they can be used for clustering.

Section 2 motivates this approach in the case of location clustering. In Section 3 it is applied to clustering regression data in the case of classical vertical regression (Section 3.1) and in the case of orthogonal regression, which has several advantages (Section 3.2). In Section 4 the orthogonal regression method is used for identifying edges in noisy images.

All proofs can be found in Müller and Garlipp (2003).

2 Clustering with redescending M-estimators

Let $y_N = (y_{1N}, \ldots, y_{NN})$ be a realization of independently and identically distributed random variables $Y_{nN} \in \mathbb{R}^k$ following a distribution with density

h. The positions of the local maxima of the density h are considered as true cluster center points and are denoted by $\mathcal{M}$, i.e.

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

If the distribution of Y_{nN} is a mixture of distributions with unimodal densities, for example $Y_{nN} = \mu_l + E_{nN}$ with probability γ_l ($\sum \gamma_l = 1$) and E_{nN} has density f_l with maximum at 0, then the local maxima of the density h are attained at the maxima μ_l of the densities $f_l(\cdot - \mu_l)$ only if the supports of the $f_l(\cdot - \mu_l)$ do not overlap, what in general is not the case. Nevertheless, to define the true cluster center points via the maxima of the density h is more general, since the densities within the clusters are not known in practice and this definition is even appropriate for the general situation, where no assumptions for the model are made and only a general density h is used. Hence the 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)), we estimate the local maxima of h and thus the center points by the local maxima of the estimated density given by the kernel estimator. In Theorem 1 we show the consitency of this estimator under some regularity conditions. A kernel density estimator for $h(\mu)$ is given by

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

where $\mu \in \mathbb{R}^k$, $\rho : \mathbb{R}^k \to \mathbb{R}^+$ is the kernel function and $s_N \in \mathbb{R}^+ \setminus \{0\}$ is the bandwidth. If s_N converges to zero, then $H_N(\mu, y_N)$ converges to $h(\mu)$ in probability under some regularity conditions. Hence, the local maxima of $H_N(\cdot, y_N)$, which also can be considered as M-estimates with respect to the objective function $H_N(\cdot, y_N)$, can be used as estimate for the set $\mathcal{M}$.

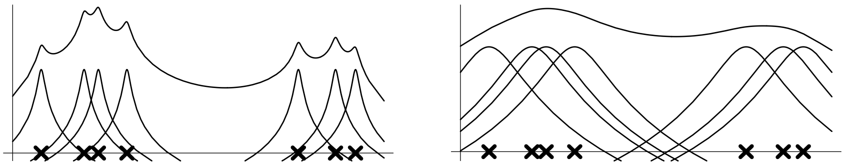

Fig. 1. Some one dimensional observations with corresponding score functions and their sum (objective function) with small (left) and large (right) scale parameter.

Usually ρ will be a unimodal density. Hence, if the scale parameter s_N is small enough and the distance between the y_{nN} are large enough, every y_{nN} is a local maximum. But usually there is so much overlap of the $\rho\left(\frac{1}{s_N}(y_{nN} - \mu)\right)$ that none of the y_{nN} is a local maximum (Figure 1). However, searching the local maxima in increasing direction starting at any y_{nN}

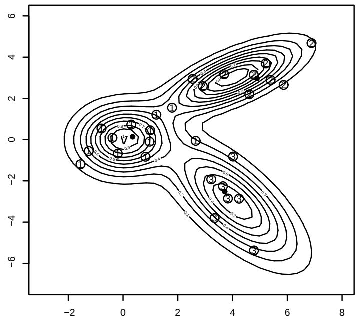

Fig. 2. Contour plot of the density of the mixture of three two dimensional normal distributions with generated observations and estimated cluster center points.

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 for the consistency of the set

$$M_N(y_N) := \{\mu \in \mathbb{R}^k;\ H_N(\mu, y_N) \text{ has local maximum at } \mu\}$$

which is the estimate of the set $\mathcal{M}$ of the positions of the true local maxima. The local maxima of $H_N(\cdot, y_N)$ can be found by Newton Raphson method starting at any y_{nN} with $n = 1, \ldots, N$.

To avoid problems like "bump huntig" (see e.g. Donoho (1988)), we need not only pointwise convergence of $H_N(\mu, y_N)$ to $h(\mu)$ but additional assumptions to achive the consistency of the set $\mathcal{M}_N(y_N)$ for the set $\mathcal{M}$. One is the uniform convergence which can be achieved by intersecting $\mathcal{M}_N(y_N)$ with a compact subset of $\mathbb{R}^k$. Appropriate compact subsets are given by

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

Then, with $\lambda_{\max} h''(\mu)$ denoting the maximum eigenvalue of $h''(\mu)$, we have

Theorem 1. *If* $\min\{|\lambda_{\max} h''(\mu)|; \mu \in \mathcal{M}_0\} > 0$*, then there exists* $\eta_0 \in \mathbb{N}$ *so that for all* $\eta \geq \eta_0, \epsilon > 0, \delta > 0$ *there exists an* $N_0 \in \mathbb{N}$ *with*

$$P\Big(\mathcal{M}_N(Y_N) \cap \Theta_\eta \subset \mathcal{U}_\delta(\mathcal{M}) \text{ and } \mathcal{M} \subset \mathcal{U}_\delta(\mathcal{M}_N(Y_N) \cap \Theta_\eta)\Big) > 1 - \epsilon$$

for all $N \geq N_0$*, where* $\mathcal{U}_\delta(\mathcal{M}) := \{\mu \in \mathbb{R}^k;$ *there exists a* $\mu_0 \in \mathcal{M}$ *with* $\|\mu - \mu_0\| < \delta\}$ *and* $\mathcal{M}_0 := \{\mu \in \mathbb{R}^k; h'(\mu) = 0$ *and* $h(\mu) > 0\}$.

Figure 2 shows a contour plot of the density $h(\mu)$ of a mixture of three two dimensional normal distributions with parameters $\mu_1 = (0,0)^\top$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\mu_2 = (4,3)^\top$, $\Sigma_2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, $\mu_3 = (4,-3)^\top$, $\Sigma_3 = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}$ and $\gamma_1 = \gamma_2 = 0.36, \gamma_3 = 0.28$ with 28 generated obsertations and the three estimated local maxima (black dots).

3 Clustering of regression data

3.1 Vertical regression

Regard a mixture of L regression models with different parameter vectors β_l. Then we have observations $z_N = (z_{1N}, \dots, z_{NN})$,which are realizations of independently and identically distributed random variables $Z_{nN} := (X_{nN}^\top, Y_{nN})^\top$, with

$$Y_{nN} = X_{nN}^\top \beta_l + E_{nN}$$

if the n'th observation is coming from the l'th cluster.

In the case of $L = 1$, the M-estimator for the regression parameter β is defined as a maximum point of the objective function

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

where $\rho : \mathbb{R} \to \mathbb{R}^+$ is the score function and $s_N \in \mathbb{R}^+ \setminus \{0\}$ is a scale parameter (see e.g. Huber (1973, 1981), Hampel et al. (1986)).

If ρ is not convex, what means that the derivative of ρ is redescending, then $H_N(\cdot, z_N)$ has several local maxima. As Morgenthaler (1990), Hennig (1997, 2003), and Chen et al. (2001) already proposed, these can be used for finding regression clusters. Under some regularity conditions for $s_N \to 0$ and ρ, we have then

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

in probability for all $\beta \in \mathbb{R}^p$, where G_l is the distribution of X_{nN} coming from the l'th cluster and f denotes the density function of the distribution of E_{nN}. Again $\gamma_l > 0$ denotes the probability that the n'th observation is coming from the l'th cluster and $\sum_{l=1}^{L} \gamma_l = 1$ holds. The function h plays now the same role as the density h in multivariate density estimation.

Under enough separation the local maxima of h are attained at $\beta_1, \dots, \beta_L$. Hence as in the multivariate case we 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.

$$\mathcal{M} := \{\beta \in \mathbb{R}^p;\ h(\beta) \text{ has local maximum at } \beta\},$$

which can be estimated by

$$M_N(z_N) := \{\beta \in \mathbb{R}^p;\ H_N(\beta, z_N) \text{ has local maximum at } \beta\}.$$

The local maxima of $H_N(\cdot, z_N)$ can be found by Newton Raphson method starting at any hyperplane through $(x_{n_1N}^\top, y_{n_1N}), \dots, (x_{n_pN}^\top, y_{n_pN})$ with $\{n_1, \dots, n_p\} \subset \{1, \dots N\}$.

As in the multivariate case, $\mathcal{M}_N(z_N)$ is a consistent estimator for $\mathcal{M}$ if it is intersected with a compact subset, which is here

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

However, here the compactness of Θ_η is not always satisfied. In particular, 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.

Hence, with $\mathcal{U}_\delta(\mathcal{M})$ and $\mathcal{M}_0$ as in Theorem 1 we have the

Theorem 2. *If Θ_η is compact for all $\eta \in \mathbb{N}$ and $\min\{|\lambda_{\max} h''(\beta)|; \beta \in \mathcal{M}_0\} > 0$, then there exists $\eta_0 \in \mathbb{N}$ so that for all $\eta \geq \eta_0, \epsilon > 0, \delta > 0$ there exists an $N_0 \in \mathbb{N}$ with*

$$P\Big(\mathcal{M}_N(Z_N) \cap \Theta_\eta \subset \mathcal{U}_\delta(\mathcal{M}) \text{ and } \mathcal{M} \subset \mathcal{U}_\delta(\mathcal{M}_N(Z_N) \cap \Theta_\eta)\Big) > 1 - \epsilon$$

for all $N \geq N_0$.

3.2 Orthogonal regression

For orthogonal regression usually an error-in-variable model is assumed. Considering a mixture of L regressions with parameters $(a_l^\top, b_l) \in S_1 \times \mathbb{R}$, $(S_1 = \{a \in \mathbb{R}^p : ||a|| = 1\})$, this means that we have observations $z_N = (z_{1N}, \ldots, z_{NN})$, which are realizations of independent and identically distributed random variables $Z_{nN} := (V_{nN}^\top, W_{nN})^\top$, 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}, $E_{1nN} \in \mathbb{R}^{p-1}$, Y_{nN}, W_{nN}, $E_{2nN} \in \mathbb{R}$, and

$$a_l^\top \begin{pmatrix} X_{nN} \\ Y_{nN} \end{pmatrix} = b_l \text{ almost surely,}$$

for Z_{nN} coming from the l-th regression.

In the case of $L = 1$, an M-estimator for (a, b) was proposed by Zamar (1989) and extends the orthogonal least squares regression estimator. It is defined as a maximum point of the objective function

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

where $\rho : \mathbb{R} \to \mathbb{R}^+$ is the score function and $s_N \in \mathbb{R}^+ \setminus \{0\}$ is a scale parameter.

For finding regression clusters, redescending M-estimators for orthogonal regression were also proposed by Chen et al. (2001). Under some regularity conditions for $s_N \to 0$ and ρ, we have then

$$H_N(a, b, Z_N) \overset{N\to\infty}{\longrightarrow} h(a, b)$$

in probability for all $(a^\top, b) \in S_1 \times \mathbb{R}$, where $h(a, b) = f_{a^\top Z_{nN}}(b)$ is the density

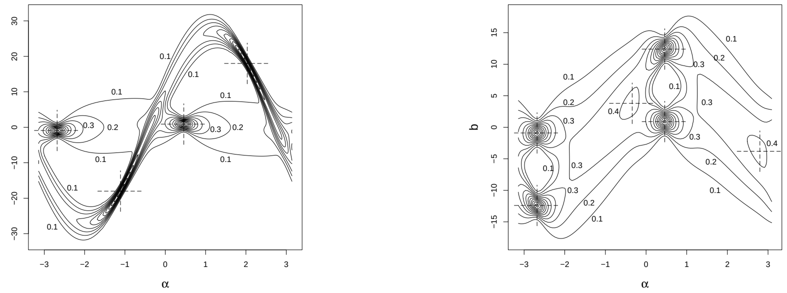

Fig. 3. Contour plot of the limit function $h(a, b)$ for a mixture of two nonparallel (left, $(\alpha_1, b_1) = (0.46, 0.9)$, $(\alpha_2, b_2) = (-1.11, -18)$) and two parallel (right, $(\alpha_1, b_1) = (0.46, 0.9)$, $(\alpha_2, b_2) = (0.46, 12.4)$) regression lines.

of the distribution on $a^\top Z_{nN}$. Note that, in opposite to classical vertical regression, the function $h(a, b)$ again is a density and shows therefore more relations to the function h in the multivariate case of Section 2. Moreover, as in Section 3.1, h is independent of ρ.

If the regression hyperplanes given by $(a_l^\top, b_l)$ are enough separated, then $h(a, b)$ will have local maxima at $(a, b) = (a_l, b_l)$.

See for example Figure 3 for the two-dimensional case with $a_l = (\cos(\alpha_l), \sin(\alpha_l))^\top$. Note that the symmetry in Figure 3 is caused by the π-periodicity of the parameter α. 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. Figure 4 shows a simulation for both cases. Here, with the used scale parameter $s_N = 2$, the objective function $H_n(a, b, z_n)$ has a third local maximum also in the case of nonparallel lines but again with a smaller height.

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

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

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

$$M_N(z_N) := \\ \{(a^\top, b) \in S_1 \times \mathbb{R};\ H_N(a, b, z_N) \text{ has local maximum at} (a^\top, b)\}. \tag{1}$$

The local maxima of $H_N(\cdot, z_N)$ can be found as for vertical regression (see Section 3.1).

As before, the consistency of $\mathcal{M}_N(z_N)$ can be shown only if $\mathcal{M}_N(z_N)$ is intersected with the set

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

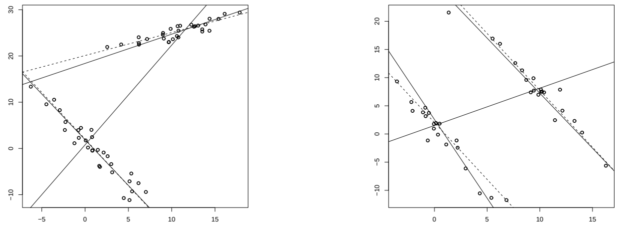

Fig. 4. True (dashed) and estimated (solid) regression lines in the case of two nonparallel and two parallel regression lines.

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 of the regressor X_{nN}. Hence, orthogonal regression is also in this sense superior to classical vertical regression where a restriction on the distribution of X_{nN} is necessary to ensure the compactness of Θ_η (see Section 3.1).

With $\mathcal{U}_\delta(\mathcal{M})$ and $\mathcal{M}_0$ as in Theorem 1 we have the

Theorem 3. *If* $\min\{|\lambda_{\max} h''(a,b)|; (a^\top, b) \in \mathcal{M}_0\} > 0$*, then there exists* $\eta_0 \in \mathbb{N}$ *so that for all* $\eta \geq \eta_0, \epsilon > 0, \delta > 0$ *there exists an* $N_0 \in \mathbb{N}$ *with*

$$P\Big(\mathcal{M}_N(Z_N) \cap \Theta_\eta \subset \mathcal{U}_\delta(\mathcal{M}) \text{ and } \mathcal{M} \subset \mathcal{U}_\delta(\mathcal{M}_N(Z_N) \cap \Theta_\eta)\Big) > 1 - \epsilon$$

for all $N \geq N_0$*.*

4 Edge identification

As an application of the orthogonal regression cluster method, we use it to detect edges in noisy images (see Figure 5.A). 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 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 at 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 (1) in Section 3.2). Applying a multiple test based on the RDKE-method, we found 2199 points, which could belong to one of the edges (see Figure 5.B). For details see Müller and Garlipp (2003). On these points, we applicate the orthogonal regression estimator with the density of the standard normal distribution as score function ρ. The scale parameter s_N is choosen with respect to the window size of the RDKE method (for details, see again Müller and Garlipp (2003)). For deciding which of the seven found center lines (Figure 5.C) belong to the true clusters, we used the absolute height of the local maxima. The result is shown in Figure 5.D.

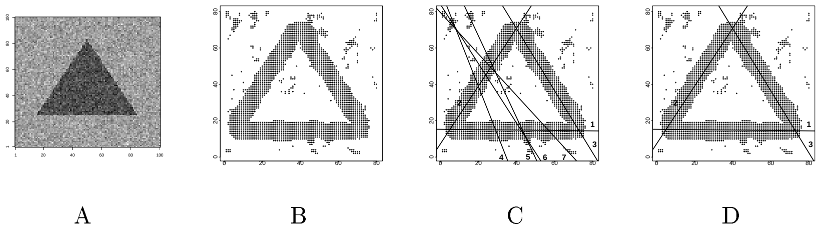

Fig. 5. Original image with 100 × 100 pixels, overlayed by normal distributed noise (A); Estimated jump points, respectively observations $z_{n,2199}$ (B); Observations z_{2199} with the estimated cluster lines $M_{2199}(z_{2199})$ (C); Observations with the three center lines with the largest maxima (D).

References

CHEN, H., MEER, P., TYLER, D.E. (2001): Robust regression for data with multiple structures. *Computer Vision and Pattern Recognition Conference, Kauai, Hawaii, December 2001, vol. I, 1069-1075.*

CHU, C.K., GLAD, I.K., GODTLIEBSEN, F. and MARRON, J.S. (1998): Edge-preserving smoothers for image processing. *Journal of The American Statistical Association, 93, 526-541.*

DONOHO, D.L. (1988): One-sided inference about functionals of a density. *Annals of Statistics, 16, 1390-1420.*

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.

HENNIG, C. (1997): Fixed Point Clusters and Their Relation to Stochastic Models. In: Klar, R. and Opitz, O. (Eds.): *Classification and knowledge organisation.* Springer, Berlin, 20-28.

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

HILLEBRAND, M. and MÜLLER, CH.H. (2001): On consistency of redescending M-kernel smoothers. *Submitted.*

HUBER, P.J. (1973): Robust regression: Asymptotics, conjectures, and Monte Carlo. *Annals of Statistics, 1, 799-821.*

HUBER, P.J. (1981): *Robust Statistics.* John Wiley, New York.

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

MÜLLER, CH.H and GARLIPP, T. (2003): Simple consistent cluster methods based on redescending M-estimators with an application to edge identification in images. *Revised version for: Journal of Multivariate Analysis.*

QIU, P. (1997): Nonparametric estimation of jump surface. *The Indian Journal of Statistics, 59, Series A, 268-294.*

SILVERMAN, B.W. (1986): *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.

ZAMAR, R. H. (1989): Robust estimation in the errors-in-variables model. *Biometrika, 76, 149-160.*