The Fitting of Multifunctions: An Approach to Nonparametric Multimodal Regression

Jochen Einbeck[1] and Gerhard Tutz[2]

[1] Department of Mathematics, National University of Ireland, Galway, Ireland.
[2] Institut für Statistik, Ludwig Maximilians Universität, 80799 München, Germany.

Summary. In the last decades a lot of research has been devoted to smoothing in the sense of nonparametric regression. However, this work has nearly exclusively concentrated on fitting regression *functions*. When the conditional distribution of $y|x$ is multimodal, the assumption of a functional relationship $y = m(x) + noise$ might be too restrictive. We introduce a nonparametric approach to fit *multifunctions*, allowing to assign a set of output values to a given x. The concept is based on conditional mean shift, which is an easily implemented tool to detect the local maxima of a conditional density function. The methodology is illustrated by environmental data examples.

Key words: Multi-valued regression, smoothing, conditional densities, conditional modes

1 Introduction

A typical definition of 'Nonparametric regression' is the following [5]:

"Given observations from an explanatory variable X and a response variable Y, construct a ***function****, a "smoother", which at point x estimates the average value of Y given that X = x. "*

Specifically, given a set of i.i.d. random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ sampled from a population $(X, Y) \in \mathbb{R}^2$ with joint density $f(x, y)$, one usually assumes a model of the type $Y = m(X) + \epsilon$, with some noise ϵ. Thereby $m : \mathbb{R} \longrightarrow \mathbb{R}$ is a smooth function relating X and Y in a suitable way, which may be generally expressed as

$$m(x) = \Omega(Y|X = x). \tag{1}$$

The choice of the operator $\Omega(\cdot)$ is quite crucial. The most popular settings are the expectation $\Omega(\cdot) = E(\cdot)$ or the median $\Omega(\cdot) = \mathrm{Med}(\cdot)$. Nonparametric regression in the sense of mean or median smoothing has been maturely

treated in the last decades. However, as already indicated by the definition given above, these techniques are restricted to the assumption of a *functional* relationship between predictor and response. When the conditional distribution of $Y|X$ is multimodal, this simple functional model may not adequately capture the essential relation between predictor and response, and the application of a mean or median smoother might blur important features of the data.

Another candidate for $\Omega(\cdot)$ is the mode operator. Modal regression has been proposed by Scott [8] and others, but has yet not been elaborated to construct a nonparametric multi-valued smoothing routine. It is the intention of this paper to fill this gap.

The mode differs from the mean and the median in one important aspect. While the conditional mean and median always represent a single value, a conditional density function can have several conditional maxima, which may be interpreted as *local modes*, being defined by

$$\text{local Mode}(Y|X=x) = \arg\max_{a\in U} f_{Y|X}(a|x)$$

where U (in the unidimensional case) is a closed interval and the maximum is taken from the interior of the interval. When the conditional distribution of the data is multimodal, then the data cannot be described properly by a function. Therefore it is assumed that the underlying relation $R \subset \mathbb{R}^2$ decomposes into several (almost everywhere smooth) branches, which are defined by the operators

$$\Omega_{(j)}(\cdot) = j^{\text{th}}\,\text{local Mode}(\cdot),$$

where $j = 1, \ldots, p$ is a suitable enumeration of the branches (e.g. from bottom to top). The underlying relation has the form

$$R = \{(x, \Omega_{(j)}(Y|X=x)); x \in \mathbb{R}, j = 1, \ldots, p\},$$

and the counterpart to model (1) is given by the *multifunction*

$$M(x) = \{\Omega_{(j)}(Y|X=x) | 1 \le j \le p\}.$$

The rest of this paper is organized as follows. Section 2 introduces our approach to estimate conditional modes, which is based on a simple conditional mean shift procedure. Section 3 gives some real data examples. Section 4 treats the evaluation of the relevance of the estimated branches, and the paper finishes with a conclusion in Section 5.

2 Conditional modes and densities

According to Samanta & Thavaneswaran [7], the conditional density $f(y|x) = f(x,y)/f(x)$ can be estimated by

$$\hat{f}(y|x) = \frac{1}{h_2}\sum_{i=1}^{n} w_i(x) K_2\left(\frac{Y_i - y}{h_2}\right),$$

where

$$w_i(x) = K_1\left(\tfrac{X_i - x}{h_1}\right) / \sum_{j=1}^{n} K_1\left(\tfrac{X_j - x}{h_1}\right). \tag{2}$$

For a given x, they show that the maximizer

$$y_m(x) = \arg\max_y \hat{f}(y|x),$$

called *sample conditional mode*, is a consistent and asymptotically normally distributed estimator for the conditional mode under some regularity conditions. However, it remains the problem of *how* to find the maxima of the conditional density estimates. A grid search would be on principle possible, but is computationally demanding and is not straightforwardly implemented when *all local* conditional modes (rather than the global one) are required. Thus, we will not pursue this idea further, but use a simpler, faster, and more elegant procedure. Let us assume that K_2 belongs to a special class of radially symmetric kernel functions satisfying $K_2(\cdot) = c_k k((\cdot)^2)$, with c_k being a strictly positive constant. The function $k(\cdot)$ is called the *profile* of K_2. By considering

$$\frac{\partial \hat{f}(y|x)}{\partial y} = \frac{2c_k}{h_2^3} \sum_{i=1}^{n} w_i(x) k'\left(\left(\frac{Y_i - y}{h_2}\right)^2\right)(y - Y_i)$$

and setting this expression to zero one obtains for the mode estimator y_m the equation

$$y_m = \frac{\sum_{i=1}^{n} w_i(x) k'\left(\left(\frac{Y_i - y_m}{h_2}\right)^2\right) Y_i}{\sum_{i=1}^{n} w_i(x) k'\left(\left(\frac{Y_i - y_m}{h_2}\right)^2\right)}.$$

Note that the dependence of $y_m \equiv y_m(x)$ on x is suppressed for notational ease. Let $g(\cdot) = -k'(\cdot)$ and consider g as a kernel profile belonging to a kernel function $G(\cdot) = c_g g((\cdot)^2)$. When K_2 is the Gaussian kernel, then G is Gaussian as well. By use of G and of the weight function (2) one obtains

$$y_m = \frac{\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right) G\left(\frac{Y_i - y_m}{h_2}\right) Y_i}{\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right) G\left(\frac{Y_i - y_m}{h_2}\right)}. \tag{3}$$

This equation cannot be solved analytically, but the solution y_m can be obtained iteratively by calculating a series of local means. Let $\mu(y_m)$ denote the right side of equation (3). An important tool is the so-called *mean shift* $\mu(y) - y$, which for a mode y_m takes the value zero. For a given starting point y_0, Comaniciu & Meer [2] show that the sequence $(y_\ell)_{\ell=0,1,2,\ldots}$ defined by

$$y_{\ell+1} = \mu(y_\ell) \tag{4}$$

converges to a nearby mode y_m, which is a fixed point of (4). To account for multimodal conditional distributions, one applies the mean shift procedure as follows:

Algorithm: Nonparametric multi-valued regression.

For a given x,

1. Choose a set of starting points $y_0^{(1)}(x) < \ldots < y_0^{(P)}(x)$.
2. For $j = 1, \ldots, P$: Set $\ell = 0$. Iterate
$$y_{\ell+1}^{(j)}(x) = \mu(y_\ell^{(j)}(x))$$
until convergence is reached, resulting in estimates $\hat{y}_m^{(1)}(x), \ldots, \hat{y}_m^{(P)}(x)$
3. The estimator for $M(x)$ is the random set
$$\hat{M}(x) = \{\hat{y}_m^{(1)}(x), \ldots, \hat{y}_m^{(P)}(x)\},$$
where the values $\hat{y}_m^{(j)}(x)$ are not necessarily distinct.

Note that the algorithm does not require do calculate the conditional densities themselves. The set $\hat{M}(x)$ is ordered, i.e. $\hat{y}_m^{(1)}(x) \leq \ldots \leq \hat{y}_m^{(P)}(x)$. This follows immediately from the properties of the mean shift, as the series of local means converges to a nearby conditional mode ([2], Theorem 1). This ordering makes it easy to identify the branches.

The choice of the number P of starting points depends on the number p of branches one expects. To be certain that all modes are discovered, one has to install a sufficiently large number $P \geq p$ of starting points. Each point gives an iteration process, which will find a conditional mode within its basin of attraction. The choice $P > p$ certainly implies that some branches will be found more than once, but for a sufficiently high number of iterations (usually, about 30 is enough) all estimates belonging to the same branch will be approximately equal. If one may assume that the data are bimodal, it is sufficient to start one mean shift procedure from the bottom and one from the top of the data cloud.

3 Examples

Firstly, we consider a speed-flow diagram as frequently used in transportation engineering (see e.g. [3]) for a Californian uninterrupted highway ("freeway") having 4 lanes, where only the lane 2 is considered here (Fig. 1, data from University of Berkeley). The speed is measured in miles per hour, and the flow in vehicles per lane per hour. Each point represents an average speed and hourly flow rate for data collected over a 30-seconds interval. For uncongested traffic, there is no significant association between traffic flow and speed - this is the big cluster at the top. When the traffic gets too dense, however, speed may be considerably diminished due to congestion, yielding the less dense data points at the bottom.

Looking at Fig. 1, one notices that the speed v cannot be described as a function $v(q)$ of the flow q. Thus, any attempt on modelling data of this type has been based on modelling the traffic flow as a function $q(v)$. However, traffic speed prediction (which is of interest e.g. to construct Intelligent Transportation Systems, ITS) would require exactly the opposite setting, i.e. $v = v(q)$. Fig. 1 (bottom) shows the results of a multimodal regression according to the presented algorithm, using Gaussian kernels with bandwidths $h_1 = 100$ and $h_2 = 4$. The starting points are chosen constant w.r.t. x, i.e. $y_0^{(1)}(x) \equiv y_0^{(1)} = \min\{Y_1, \ldots, Y_n\}$ and $y_0^{(2)}(x) \equiv y_0^{(2)} = \max\{Y_1, \ldots, Y_n\}$. The estimated curve is superior to the estimates based on a local mean or the local median (Fig. 1 top), which do not take account for the data points in the bottom of the plot, which obviously carry some information and cannot be discarded.

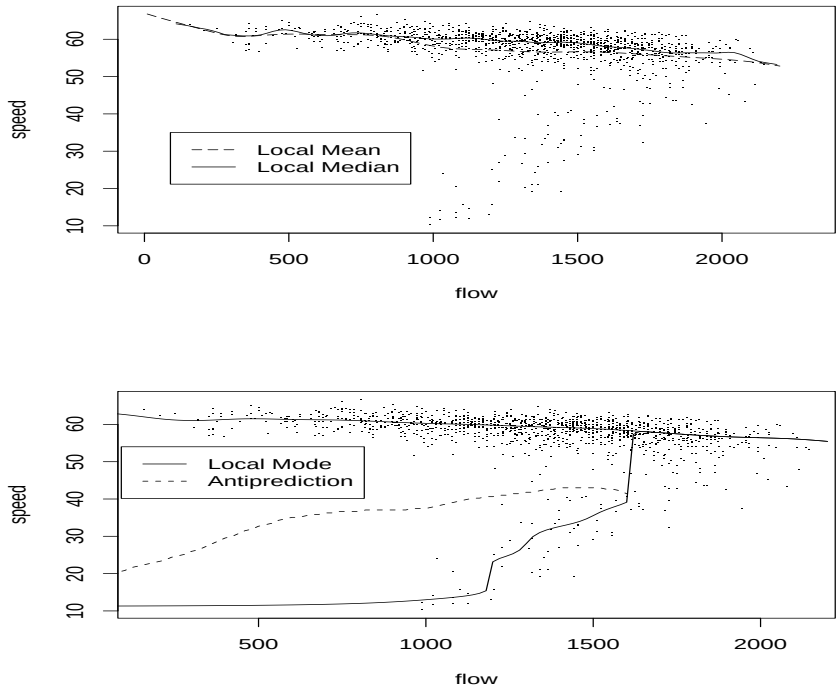

Fig. 1. Speed-flow diagram for lane 2 with local smoothers based on the conditional mean, median (top) and mode (bottom). In the bottom also the antiprediction curve (see Section 4) is plotted.

Secondly, we consider a data set which does not seem to be a candidate for the presented procedure at the first glance. The Old Faithful Geyser data (data set *faithful* in R package *datasets*), describing the waiting time (in minutes) between eruptions and the duration of the eruption for the Old Faithful geyser

in Yellowstone National Park, Wyoming, USA, have been frequently used to illustrate the performance of smoothers or density estimators; see e.g. [6]. The data and a local mean smoother (loess) are shown in Fig. 2.

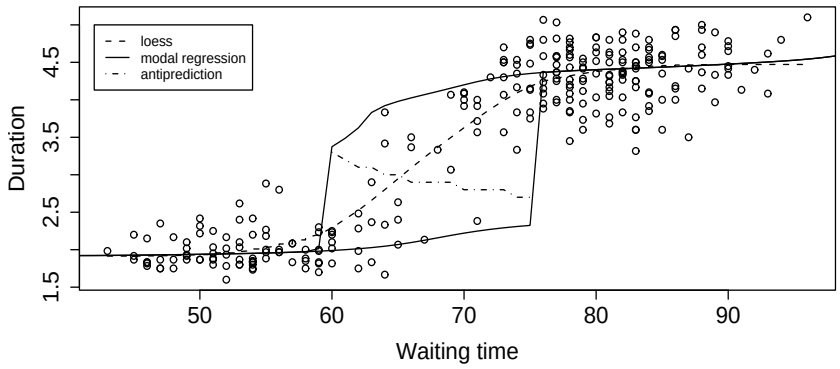

Fig. 2. Old Faithful Geyser data with loess smoother, modal regression and antiprediction curve.

Though the loess smoother seems to do a good job, it raises some problems, as also observed by Hennig [4] considering another version of the Old Faithful data. For instance, given a waiting time of $x = 68$, the loess prediction would be $\hat{y} = 3.37$. Regarding the plot in more depth, one notices that this value is in fact very unlikely. There is hardly any observed eruption duration in the interval from about 2.5 to 3.5 minutes. However, it seems to be appropriate to assume that there are two regimes, one with low waiting times and low eruption durations, and other one with higher ones, where a certain overlap between these regimes is likely. A modal regression applying the presented procedure ($h_1 = 5, h_2 = 0.27$) yields the solid lines in Fig. 2, which unveil the two-regime-structure of the data set.

4 Relevance, Antiprediction and Classification

A crucial point is the evaluation of the relevance of a conditional mode. Intuitively, the probability mass between the neighboring valleys surrounding the mode is a useful measure for the relevance of a mode. Fig. 3 (left) illustrates this concept for the speed-flow data given a flow of 1300 vehicles/hour. The area between the left border and the valley contains an estimated probability of 0.072, and the second mode corresponds to the probability 0.928. Thus, one would infer here

$$\hat{M}(1300) = \begin{cases} 28.06 & \text{with estimated prob.} \quad 0.072 \\ 59.44 & \text{with estimated prob.} \quad 0.928 \end{cases}$$

To estimate these probabilities, one has to find the lows of the valleys and to integrate over the estimated conditional densities between them. Without

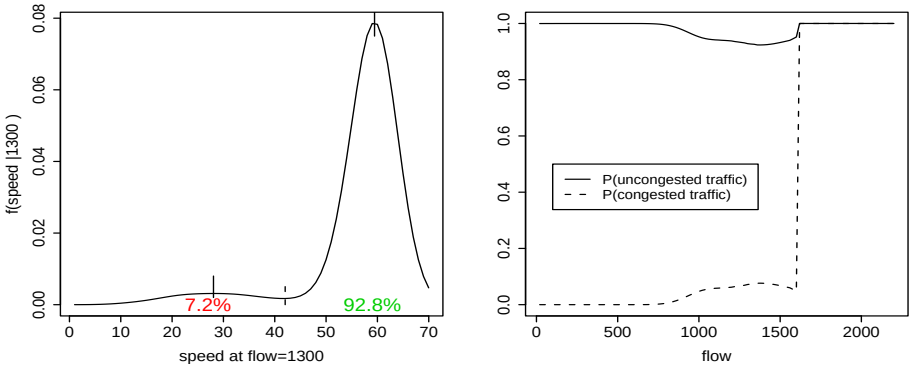

Fig. 3. Left: Estimated conditional density at a flow of 1300 vehicles/hour. Modes and antimodes are indicated by short solid and dashed vertical lines, respectively. Right: Probabilities of the branches of smooth multimodal regression curves in dependence of flow.

too much effort one can do the search for the minimum and the integration simultaneously. For a given (local) mode y_m at x, one descends from the (local) maximum $f(y_m|x)$ in small steps of length δ, say, to the right (steps $k = 0, 1, 2, \ldots$) as well as to the left (steps $k = -1, -2, \ldots$), and augments the integral in each step by $\delta \cdot f(y_m + k\delta|x)$ until the minimum is reached, i.e the sequence $(f(y_m + k\delta|x))_k$ stops to fall. Note that the number of steps until the next minimum to the left and to the right do not need to be the same. This integral is usually surprisingly accurate, as the approximation errors on the left and on the right side of the maximum tend to cancel out. The choice of δ is not very crucial, because it is not a tuning parameter, but only influences the accuracy of the approximation. Fig. 3 (right) shows the probabilities obtained in this manner for the speed-flow data. At a flow of 1620 vehicles/hour, the dashed line rises rapidly and merges with the solid one, as the components are no longer separated beyond this point. This is certainly *not* a sign for a suddenly rising probability of congested traffic.

If one stores, for a given value x, the positions of the minima found while calculating the above integrals, one obtains a vector of *conditional antimodes*. An antimode can be seen as an antiprediction - a value that is likely not to be seen. Connecting the conditional antimodes in x-direction, one obtains a nonparametric *antiregression* or *antiprediction curve*, i.e. a curve describing where the data are *not* to be expected.

In the case of speed-flow data, one observes that this curve (Fig. 1 bottom) is useful to classify the data into observations coming from the congested or uncongested regime, as long as a division is possible. The antiprediction curve for the Geyser data is plotted in Fig. 2 (dashed-dotted line) and classifies the data into regimes with high and low eruption duration.

5 Conclusion

We showed that the consideration of the conditional mode rather than the conditional mean or median is useful when the data can be assumed to be associated to several, possibly in x- and y- direction overlapping, regimes. We demonstrated how smooth modal regression curves can be calculated easily by means of a conditional mean shift procedure. We also showed how modal regression may be used to identify areas of transition between regimes.

Though being a simple 2-dimensional problem, the task of multi-valued regression has received little attention yet. However, it should be noted that there exist approaches to *parametric* multi-valued [10] and cluster-wise [4] regression, and some related contributions from computer scientists in the context of inverse mapping problems [1]. Further, there exist nonparametric multi-valued regression methods for the simpler case that it is *known a priori* which point belongs to which branch [9]. We remark finally that the extension of the presented method to multivariate predictors is straightforward by employing multivariate kernels in (2).

Acknowledgement. This work was partly supported by Science Foundation Ireland Basic Research Grant 04/BR/ M0051. Support from Deutsche Forschungsgemeinschaft (SFB 386) in various aspects is gratefully acknowledged.

References

[1] CARREIRA-PERPIÑAN, M. A. and WILLIAMS, C.K.I. (2003): On the number of modes of a Gaussian mixture. Lecture Notes in Comp. Science, 2695, 625–640.
[2] COMANICIU, D., and MEER, P. (2002): Mean Shift: A robust approach towards feature space analysis. IEEE Trans. Pattern Anal. Machine Intell., 24, 603–619.
[3] HALL, F. L., HURDLE, V. F. and BANKS, J. M. (1992): Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relations on freeways. Transportation Research Record, 1365, 12–17.
[4] HENNIG, C. (2003): Clusters, outliers, and regression: fixed point clusters. Journal of Multivariate Analysis, 86, 183–212.
[5] HOLMSTRÖM, L. (2003): New Modeling and Data Analysis Methods for Satellite Based Forest Inventory. Final Seminar, Antares Programme, www.tekes.fi.
[6] LOADER, C. (1999): Local Regression and Likelihood. Springer, New York.
[7] SAMANTA, M. and THAVANESWARAN, A. (1990): Non-parametric estimation of the conditional mode. Comm. Stat. Theory Meth., 19, 4515–4524.
[8] SCOTT, D.W. (1992): Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley, New York
[9] SILVERMAN, B. W. and WOOD, J. T. (1987): The nonparametric estimation of branching curves. Journal of the Amer. Statist. Assoc., 82, 551–558.
[10] WEDEL, M. and KAMAKURA, W.A. (1995): A mixture likelihood approach for generalized linear models. Journal of Classification, 12, 21–55.