Multivariate regression smoothing through the "falling net"

James Taylor[1] and Jochen Einbeck[1]

[1] Department of Mathematical Sciences, University of Durham, Durham, DH1 3LE, UK.

Abstract: We consider multivariate regression smoothing through a conditional mean shift procedure. By computing local conditional means iteratively over a set or grid of target points, at each iteration a "net" is formed which gently drifts towards the data cloud, until it settles at the conditional modes of the response distribution. The method is edge-preserving and allows for multi–valued response.

Keywords: Conditional density; modal regression; smoothing

1 Methodology

Given d-variate covariates $X_i = (X_{i1}, ..., X_{id})^T$ and scalar response values Y_i where $i = 1, ..., n$, we find the regression surface via the conditional modes of Y given $X = x$. These are determined by the conditional density function, $f(y|x)$, which can be estimated through

$$\hat{f}(y|x) = \frac{\hat{f}(x,y)}{\hat{f}(x)} = \frac{\sum_{i=1}^{n} G\left(\frac{Y_i - y}{b}\right) \prod_{j=1}^{d} K\left(\frac{X_{ij} - x_j}{h_j}\right)}{b \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{X_{ij} - x_j}{h_j}\right)}, \qquad (1)$$

where G and K are univariate (e.g. Gaussian) kernels, and the subscript j denotes the $j-$th component of the corresponding vector. The values b and h_j are bandwidth parameters to be selected. At each x there may be more than one conditional mode since $\hat{f}(y|x)$ can have several maxima. By setting $\frac{\partial \hat{f}(y|x)}{\partial y} = 0$, one obtains a conditional mode y_m (argument x omitted for ease of notation) as the solution to the estimation equation $y_m = \mu(y_m)$, with

$$\mu(y_m) = \frac{\sum_{i=1}^{n} G\left(\frac{Y_i - y_m}{b}\right) \prod_{j=1}^{d} K\left(\frac{X_{ij} - x_j}{h_j}\right) Y_i}{\sum_{i=1}^{n} G\left(\frac{Y_i - y_m}{b}\right) \prod_{j=1}^{d} K\left(\frac{X_{ij} - x_j}{h_j}\right)}. \qquad (2)$$

Since this cannot be solved analytically, we solve it iteratively using the result by Cheng (1995) that, starting from any $y_0 \in \mathbb{R}$, the mean shift

procedure $y_{\ell+1} = \mu(y_\ell)$ converges to a nearby conditional mode. In order to detect more than one mode for each x it is necessary to specify more than one starting point for the mean shift, typically two. For bivariate predictors, if y_0 is (for all x) set greater than all Y_i, the simultaneous iterative execution of the mean shift resembles visually a net falling onto the data and forming a surface. Of course, if y_0 is below rather than above all Y_i, we would talk about a "rising" net. We emphasize that the techniques proposed in this section do neither require the estimation of any density function, nor the solution of any optimization problem (such as least squares) at any stage; all computational work is carried out by the mean shift.

2 Examples

Figure 1 (left) shows data from a wheat yield trial, where latitude and longitude serve as covariates (the data are part of R package **nlme**, Pinheiro et al. (2008)). Figure 1 (right) provides the surface formed after 30 iterations of the mean shift process on the dataset. Here $h_1 = 3.18$, $h_2 = 3.18$ and $b = 5.61$ after using the bandwidth selection methods described in Section 4.

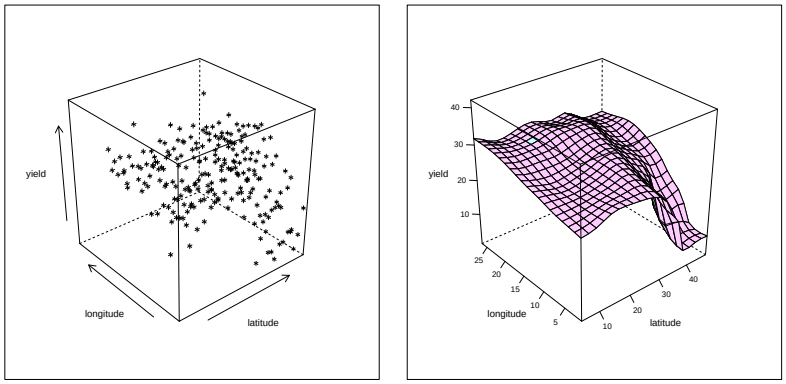

FIGURE 1. The procedure applied to the wheat yield dataset.

Figure 2 illustrates the characteristics of this smoothing technique through simulated data sets of size $n = 200$. Data set A is simulated from the univariate function $y = \sin(0.2x_1) + \cos(x_2)$ and subjected to Gaussian error with standard deviation 0.05. Data set B has a partially bimodal response, which splits for $x_1 \geq 0.5$ into two branches. For $x_1 < 0.5$ the response is simulated from the univariate function $y = 1.5 + 3x_1$ with Gaussian error of standard deviation 0.4. For $x_1 \geq 0.5$, the upper plane is centered at $y = 3$ and the lower plane at $y = 1$; the error standard deviation is 0.2

each. One observes from Figure 2 how the estimated surfaces develop after different numbers of iterations, ℓ, with starting points positioned *above* (upper estimated surface) and *below* (lower estimated surface) all responses. The right hand column of Figure 2 demonstrates clearly that the procedure is edge-preserving, and able to identify multiple branches when the underlying conditional distribution is multimodal, where other regression techniques could not successfully describe it.

3 Relevance of a mode

When there exist more than one mode of the conditional response distribution for a given x, it is interesting to evaluate the relevance of the different modes. To estimate the probability associated with a conditional mode, one integrates numerically over the part of the estimated conditional density which forms that modal peak. The search for the minimum and the integration can be done simultaneously by descending in small steps from the modes and increasing the integral until either the boundary or the next dip separating the modes is reached (Einbeck and Tutz, 2006). For the simulated data from the right hand column of Figure 2, Figure 3 (left) shows a surface of probabilities, calculated as described, showing the probability of data being present in the mode captured by the "falling net". Figure 3 (right) shows the same for the "rising net." For this data set, the plots show a probability of 1 for about half of all values of x; this is expected since the response is unimodal for these x.

4 Bandwidth selection

In the case of multivariate predictors, the problem of bandwidth selection is more challenging than in the univariate case, since values must be selected for all the h_j as well as for b. For the selection of bandwidth b, one can resort to univariate conditional density bandwidth selectors, such as *cde.bandwidths* in the package **hdrcde**, Hyndman (2010), since this bandwidth does not directly depend on d. Performing this for each covariate separately and then taking the mean as b is effective here. Given b, the h_j are successfully selected by adapting Bashtannyk and Hyndman's (2001) univariate *regression-based bandwidth selector* for use with multivariate covariates, as the authors themselves suggest doing. Therefore we standardize the covariates and search for an optimal $h = h_1 = ... = h_d$. The extended regression-based bandwidth selector minimizes the penalized average squared prediction error $Q(h)$ with respect to h, for a fixed b, where

$$Q(h) = \frac{\Delta}{n}\sum_{k=1}^{N}\sum_{i=1}^{n}\left\{\frac{1}{b}G\left(\frac{Y_i - y'_k}{b}\right) - \hat{f}(y'_k|X_i)\right\}^2 \times p\left(\frac{(K(0))^d}{\sum_{l=1}^{n}\prod_{j=1}^{d}K\left(\frac{X_{ij}-X_{lj}}{h}\right)}\right) \quad (3)$$

where $\{y'_1, ..., y'_N\}$ are equally spaced over the sample space Y with $y'_{i+1} - y'_i = \Delta$ and where $p(u) = (1-u)^{-2}$ is a penalty function. This $p(u)$ is identical to that used in generalized cross-validation, but differs from the one used typically in the univariate case for this technique, since this was found to perform badly in the multivariate setting. Once h has been found, it is unstandardized and the modal regression is then carried out with unstandardized covariates and bandwidths. Following this procedure for the wheat yield data gives the bandwidths stated in Section 2.

5 Discussion

This work constitutes essentially a multivariate extension of the multimodal regression technique introduced in the context of traffic data modelling in Einbeck and Tutz (2006). The problem of bandwidth selection has been addressed by appropriately extending bandwidth selectors which were developed for conditional density estimation with univariate predictors by Bashtannyk and Hyndman (2001).

Attractive features of the technique are the computational simplicity, the edge-preserving property, and the visual appeal. Moreover, the method is able to deal with multi–valued response, though it should be admitted that data of this type are relatively rare, and that multiple modes in the response distribution may be an indicator that important covariates have been omitted from the model. Nevertheless, the presented approach may still serve to detect and visualize situations of this type.

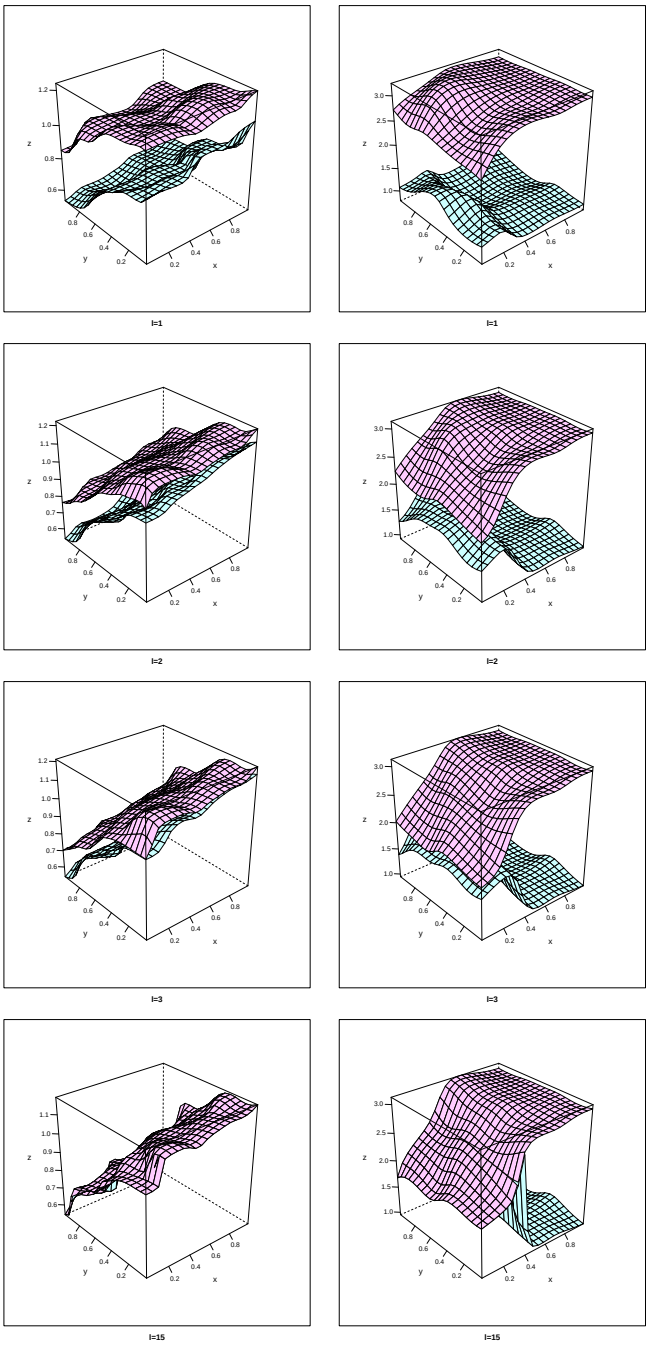

FIGURE 2. The left column displays the surfaces for simulation A, for $\ell = 1, 2, 3, 15$ (from top to bottom). The right column shows the same for simulation B.

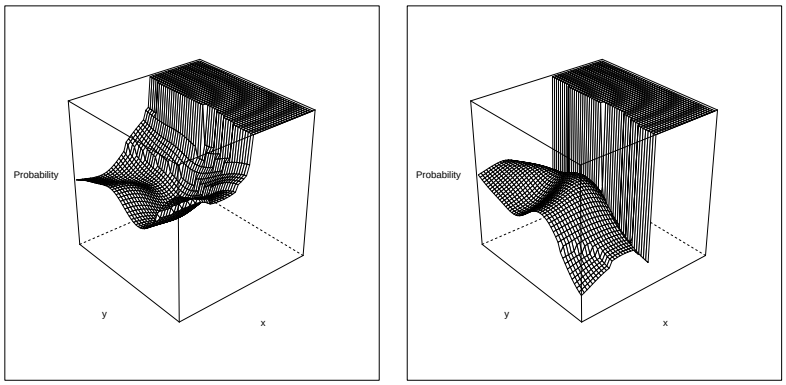

FIGURE 3. Left: Bivariate probability plot for the "falling net" (left) and the "rising net" (right); each for the fitted surface from Figure 2 (bottom right). Note that the orientation is rotated in order to allow for a better view of the probability surfaces.

References

Bashtannyk, D. and Hyndman, R. (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics and Data Analysis*, **36**, 279–298.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Machine Intell.*, **17**, 790–799.

Einbeck, J. and Tutz, G. (2006). Modelling Beyond Regression Functions: An Application of Multimodal Regression to Speed-flow Data. *Applied Statistics*, **55**, 461–475.

Hyndman, R. (2010). **hdrcde**: Highest density regions and conditional density estimation. R package version 2.15.

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D. and the R Dev. Core Team (2011). **nlme**: Linear and Nonlinear Mixed Effects Models. R package version 3.1-98.