Strategies for local smoothing in high dimensions: using density thresholds and adapted GCV

James Taylor[1] and Jochen Einbeck[1]

[1] Department of Mathematical Sciences, University of Durham, Durham, DH1 3LE, UK, {james.taylor1, jochen.einbeck}@durham.ac.uk

Abstract: Local polynomial fitting for univariate data has been widely studied and discussed, but up until now the multivariate equivalent has often been deemed impractical, due to the so-called *curse of dimensionality.* Here, rather than discounting it completely, we use density as a threshold to determine where over a data range reliable multivariate smoothing is possible, whilst accepting that in large areas it is not. An adapted version of generalized cross-validation for multivariate bandwidth selection is also discussed.

Keywords: Smoothing, Density, Threshold, Bandwidth

1 Introduction

We are given d-dimensional covariates $X_i = (X_{i1}, ..., X_{id})^T$ and response values Y_i where $i = 1, ..., n$. Local polynomial regression is a nonparametric way of estimating the mean function $m(x) = E(Y|X = x)$. Assumed is that

$$Y_i = m(X_i) + \epsilon_i \tag{1}$$

where ϵ_i are random variables with zero mean and variance σ^2. We concentrate on local linear regression where hyperplanes of the form $\beta_0 + {\beta_1}^T x$, where β_1 and x are both vectors, are fitted locally. For each point x in d-dimensional space one minimizes

$$\sum_{i=1}^{n} \left\{ Y_i - \beta_0 - \sum_{j=1}^{d} \beta_{1j}(X_{ij} - x_j) \right\}^2 K_H(X_i - x) \tag{2}$$

with respect to $\beta = (\beta_0, \beta_{11}, ..., \beta_{1d})^T$, yielding the estimator of the mean function $\hat{m}(x) = \hat{\beta}_0$. Here, $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$ where K is a multivariate kernel function and H is the bandwidth matrix. For K, we use primarily a product of Gaussian kernels since this is the least temperamental kernel function in regions where data is sparse, which occur more often in higher dimensions. H is crucial in determining the amount and direction of

smoothing, and we choose to use a diagonal matrix, $H = \text{diag}(h_1^2, ..., h_d^2)$, for computational ease. We have adjusted generalized cross-validation slightly for use with multivariate data, and this is detailed in Section 3.
The problem primarily addressed here is the *curse of dimensionality* which refers to the issues that arise when data becomes very sparse in higher dimensions. If there is not sufficient data in a neighbourhood, then the variance of the fit is too high, or with some kernel functions, such as the popular Epanechnikov kernel, the calculations just break down completely. Often, local polynomial fitting is abandoned as a result of these problems, and other methods such as the additive models suggested in Hastie and Tibshirani (1990), are favoured. Local polynomial fitting however has the big advantage of being considerably more flexible. In Section 2, we pursue a technique to avoid the curse of dimensionality, enabling us to achieve the best possible estimate of m where sufficient information is available.

2 Density as a threshold

The method is one which essentially ignores all neighbourhoods which don't contain enough data, and so only performs smoothing over some region in which estimation is considered reliable, where the bias and variance of $\hat{m}$ can be kept reasonably low. In this way the curse of dimensionality is avoided. This method is not universal in the sense that it doesn't give estimates over the whole data range, but it is satisfactory in the sense that it gives estimates, with all the advantages of local polynomial regression, in some areas. To find these areas, and to discover where there is enough data, we examine the density f of X. The density estimate for a multivariate point x is;

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K_H(x - X_i) \tag{3}$$

In calculating the density, again a bandwidth matrix is needed, and for our purposes it is advisable to use the same parameters here as in the regression, for reasons which will become clear. We seek a threshold T such that, if at point x we have $f(x)>T$, then an estimate using local linear regression can be considered somewhat reliable, and otherwise, care should be taken and an alternative method sought, possibly local constant fitting. According to Loader (1999), one has $\frac{1}{\sigma^2}\text{Var}(\hat{m}(X_i)) \leq \text{infl}(X_i) \leq 1$. Hence, bounding the influence implies bounding the variance. Using the asymptotic approximation of the influence function given in Loader (1999), a natural choice of T is straightforwardly derived from the latter inequality;

$$T = \frac{\rho K(\mathbf{0})}{n \prod_{i=1}^{d} h_i} \tag{4}$$

where

$$\rho = e_1^T \left(\int_a^\infty K(\mathbf{v}) A(\mathbf{v}) A(\mathbf{v})^T d\mathbf{v} \right)^{-1} e_1, \tag{5}$$

$\mathbf{v} = (v_1, ..., v_d)^T$ and $A(\mathbf{v}) = (1, \mathbf{v})^T$. The position of the bandwidth parameters in the denominator is justified since with larger bandwidth parameters, in areas of relative high density, the density at x will be lower than with smaller parameters, and so a lower threshold is needed.
The parameter a appearing in the lower integral limit reflects the distance to the boundary of f for which the criterion is optimized. Based on extensive testing in the local linear case we recommend the value $a = -0.85$, corresponding to a point situated $0.85h_i$ inside the boundary. This is quite intuitive as this is just about the region where one would assume data sparsity to become a problem.

3 Adapted GCV

The *curse of dimensionality* causes problems in the area of bandwidth matrix selection too. We believe in the use of a classical method over a plug-in one due to the reliance on asymptotics of the latter. The asymptotic assumption of bandwidths tending to zero seems to be inappropriate in order to select the relatively large bandwidths needed for multivariate local smoothing.
One such classical method is generalized cross-validation which is less precise than other cross-validation, but computationally less demanding. We propose an adaptation to this which, using the median and weighting, removes the influence of data points in less dense areas which otherwise may have a disproportionate effect on the procedure, and can cause extreme values of h_i being chosen. This effect is more likely to occur as d increases. The minimization of the below has been trialled with some success;

$$AGCV(H) = n^{-1} \sum_{i=1}^{n} \left\{ \frac{Y_i - \hat{m}_H(X_i)}{1 - \psi} \right\}^2 w(X_i) \tag{6}$$

where ψ is the median of the diagonal elements of the smoother matrix after excluding the elements contributed by the points for which $w(X_i) = 0$. We set $w(X_i) = 1$ for all i except the r points at which $f(X_i)$ are smallest, at which it is 0. r is the number of points which could be considered isolated i.e. where the density at that point is equal to the density of just one data point. This is best examined using Epanechnikov kernels. The bandwidth parameters to be used in the density estimation here should be the optimal values calculated from an external source such as the *np* package in R. Choosing $r > 0$ is both a matter of finetuning by focussing on the denser region in which we are interested, and also removing any computational constraint imposed by points in sparser regions.

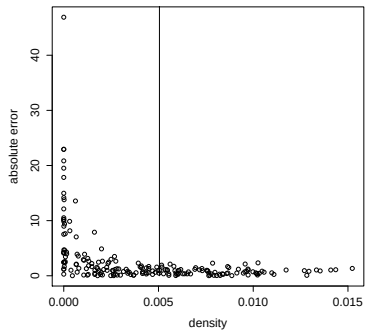

FIGURE 1. $|m(X_i) - \hat{m}(X_i)|$ v. $\hat{f}(X_i)$

4 Simulation

We simulated 3-dimensional covariates through a t-distribution with 2 degrees of freedom centered at 15.5. The response values were generated according to the model (1) with $m(X_i) = -12\cos(X_{i1}) + 5\sin(5X_{i2}) + 10\log(X_{i3}) + 17$ and $\epsilon_i \sim N(0,1), i = 1, ..., 500$. 300 of these points were used to estimate m while the remaining 200 were used to test the threshold method. In this simulation, of the 200 points tested, the threshold of 0.00505 excluded 121 of them. The value of $a = -0.85$ consistently excluded the poorest performing points, whilst deeming the better points, in terms of the absolute error, $|m(x_i) - \hat{m}(x_i)|$, apt for smoothing. This is shown in Fig.1. AGCV was also successful with this simulation. The usual GCV method suggested some h_i greater than the data range which is clearly unacceptable, but with $r = 39$, AGCV suggested a much more reasonable $H = \text{diag}(0.347^2, 0.129^2, 2.92^2)$. The density bandwidth parameters were selected using the *np* package. Simulations were performed satisfactorily with 1,3 and 16-dimensional covariates. The values of ρ calculated for use in the simulations were 1.5, 3.12 and 147.3 respectively.

References

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman and Hall.

Loader, C. (1999). *Local Regression and Likelihood.* Springer.