ELSEVIER

Statistics & Probability Letters 51 (2001) 327–336

STATISTICS & PROBABILITY LETTERS

www.elsevier.nl/locate/stapro

Variable selection by stepwise slicing in nonparametric regression

K.B. Kulasekera *

Department of Mathematical Sciences, Clemson University, Box 341907, Martin Hall, Clemson, SC 29634, USA

Received January 2000; received in revised form July 2000

Abstract

We consider variable selection issue in a nonparametric regression setting. Two stepwise procedures based on variance estimators are proposed for selecting the significant variables in a general nonparametric regression model. These procedures do not require multidimensional smoothing at intermediate steps and they are based on formal tests of hypotheses as opposed to existing methods in the literature. Asymptotic properties are examined and empirical results are given.

Keywords: Design variables; Nonparametric test; Smoothing

1. Introduction

In many studies, experimenters gather multivariate observations. One model that is used to explain such data is the multiple regression model. Specifically, if the observed data takes the form $\{Y_i, x_{1i}, \ldots, x_{pi}\}$, $i = 1, \ldots, n$, we are interested in the nonparametric model

$$Y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})'$, f is the deterministic portion of the model and ε_i is the ith noise term. We will make the typical assumptions that the ε's to be mean zero independent random variables with finite second moments. In this nonparametric model, we only assume that f has some smoothness properties such as a continuous second derivative. Due to the difficulties in estimation and interpretation, it is important to use an optimal number of covariates in the model.

The variable selection and related issues in a nonparametric setting have been examined by various authors. Li (1991) examines the dimension reduction by sliced inverse regression (SIR), where one determines some lower dimensional projections (optimal linear combinations) of the original set of covariates under certain assumptions of the error structure. Once the suitable linear combinations are found, a multivariate surface of

* Tel.: +1 864 656 5231; fax: +1 864 656 5230.
E-mail address: kk@ces.clemson.edu (K.B. Kulasekera).

0167-7152/01/$ - see front matter
PII: S0167-7152(00)00167-X

these linear combinations is fitted. Projection pursuit regression (PPR) introduced by Friedman and Stuetzle (1981) is also a way of reducing the dimension effectively. However, as many authors have noted, PPR has interpretational difficulties except possibly in the case of single index models. As an alternative to fitting general regression surfaces, Hastie and Tibshirani (1990) proposed the generalized additive model (GAM) which is somewhat a special model that does not require multivariate smoothing. However, in a GAM, one mostly gets an additive approximation to the true surface.

Selection of variables in a general multiple regression surface was examined by Zhang (1991). This method requires fitting a higher dimensional surface at every stage a new variable is considered. This induces the curse of dimensionality very early in the selection. Härdle and Korostelev (1996) discuss the selection of variables in a GAM. Their method, in spite of the simulation results that were reported, also needs multivariate density estimation at every stage, which is not a very desirable feature. In addition, the decision rule involves an arbitrary threshold parameter c_0 that has a major influence in the accuracy. Both these procedures are adhoc methods. The selection criteria do not depend on a formal test of hypothesis. The sampling properties of either method are unknown. Therefore the behavior of these selection criteria cannot be compared to each other. Also, when the variables being considered do not have a significant role in the model, the performance of these criteria is not known.

In this article, we address the variable selection in nonparametric multiple regression without higher dimensional smoothing during intermediate stages. The proposed methods have a close resemblance to the classical ANOVA method in which estimators of the variances are compared. The first method proposed here applies to models with nonadditive behavior in the candidate covariate. The second method we propose can be applied to both additive and nonadditive models equally well.

We organize the paper in the following manner. Section 2 gives the detailed procedures and the asymptotic results. Section 3 gives results of a small simulation study which examines the power properties of the proposed procedures. We also apply the selection methods to a real data set which tests the significance of a second variable in a regression model in the presence of one covariate. In Section 4, we briefly discuss various issues that arise in the implementation of the procedure such as the grouping of slices in higher dimensions and the stopping rule.

2. Selection method

In this section, we describe the selection rules and discuss their asymptotic properties. We begin with the case with two predictors; x_1 and x_2. We assume our data is of the form $\{Y_{ij}, x_{1i}, x_{2ij}\}$, $j=1,\ldots,n_i$, $i=1,\ldots,N$. This assumption can actually be relaxed to an assumption that one only requires multiple responses within a slice $[x_{1i}-\delta, x_{1i}+\delta]$ for each $\delta>0$ for a properly selected set of i's. At this point a "slice" is defined as a covariate value for the first covariate. For simplicity, we also assume that the covariates x_1 and x_2 are nonrandom and both take on values in $[0,1]$. These assumptions are readily removed with mild conditions on the design densities of the covariates. Without loss of generality, we assume that the x_{1i}'s and the x_{2ij}'s for each i are ordered. The Method 1 given here is recommended for nonadditive models while the Method 2 applies to any model. We may need some of the following assumptions at various stages.

$\mathscr{A}_1$: The sample sizes $n_i \to \infty$ for each i as $N \to \infty$.

$\mathscr{A}_2$: The function f is twice continuously differentiable in all arguments.

$\mathscr{A}_3$: The sequences of design points for each covariate become dense in $[0,1]$ as $n \to \infty$.

We define quantities

$$\chi_i^2 = \frac{\sum_{j=2}^{n_i} [Y_{ij} - Y_{i1}]^2}{2(n_i - 1)}, \quad i = 1,\ldots,N. \tag{2}$$

These can be thought of as estimators of the error variance under the null hypothesis H_0: $f(x_1, x_2) = f_0(x_1)$ which indicates no effect from variable x_2 in the model.

2.1. Method 1

Let $m = [N/2]$, where $[.]$ indicates the integer part, and combine the above variance estimators to form a test statistic

$$T = \frac{1}{\sqrt{m}} \left[\sum_{i=1}^{m} \chi_i^2 - \chi_{i+m}^2 \right]. \tag{3}$$

It is easy to see that under the null hypothesis, $E(T) = 0$. If the null hypothesis is not true and the function f is nonadditive in the covariates,

$$E(T) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \left[\frac{\sum_{j=2}^{n_i} (f(x_{1i}, x_{2ij}) - f(x_{1i}, 0))^2}{2(n_i - 1)} - \frac{\sum_{j=2}^{n_{i+m}} (f(x_{1i+m}, x_{2ij}) - f(x_{1i+m}, 0))^2}{2(n_{i+m} - 1)} \right].$$

Thus, we propose to reject the null hypothesis of no effect from x_2 for large absolute values of T. Asymptotic critical points for a test of this type can be obtained using a normal distribution due to the following theorem.

Theorem 2.1. *Under the null hypothesis and assumptions $\mathcal{A}_1$–$\mathcal{A}_3$, the statistic T converges in distribution to a normal random variable with mean 0 and variance $\tau^2 = E(\varepsilon^4) - \sigma^4/2$.*

Proof. Note that $E(\chi_i^2 - \chi_{i+m}^2) = 0$ under the null hypothesis. This, combined with the independence of χ_i^2 and χ_{i+m}^2 for each i and the fact that

$$\operatorname{var}(\chi_i^2) = \frac{(n_i - 1)E(\varepsilon^4) - n_i\sigma^4}{4(n_i - 1)}$$

enables us to use classical central limit theorem arguments to get the desired result. □

Remark 2.1. Actually what is tested is a broader null hypothesis of type

$$H_0: \int_0^t \Delta(x)\,dx - \int_t^{0.5+t} \Delta(x)\,dx + \int_{0.5+t}^1 \Delta(x)\,dx = 0,$$

with $t = 0.5$ where, $\Delta(x) = \frac{1}{2}\int_0^1 [f(x, y) - f(x, 0)]^2\,dy$. It is noted that for a large sample size,

$$\frac{E(T)}{\sqrt{N}} \approx C \left(\int_0^{1/2} \Delta(x)\,dx - \int_{1/2}^1 \Delta(x)\,dx \right).$$

Thus, with this particular grouping of the χ^2 values, there may be cases where $|E(T)| \approx 0$ even under some alternative hypotheses, especially if the function f is periodic. One does not necessarily have to form the statistic T by grouping the first and last m estimators of the variance as above. The grouping can be done in many ways, for example, one may group first fourth and the last fourth of the χ^2 values with the middle 50% ($t = 0.25$). In practice, one may do a preliminary study using several smoothing parameters in two dimensions to examine obvious patterns and determine an appropriate grouping.

Remark 2.2. Finding the best grouping for optimal power is somewhat analogous to selection of a smoothing parameter to optimize the power of nonparametric tests for comparing regression models (Kulasekera and Wang, 1997). One does not have to group the χ^2 values corresponding to x_1 values in $A = [0, t] \cup [t + 0.5, 1]$

and A^c. The grouping may be done differently as long as the particular grouping yields a test statistic T such that $|E(T)| > 0$ under the alternative hypothesis. The following lemma establishes that for all nonconstant functions in the plane, one can always find a grouping (and a test statistic T for which $|E(T)| > 0$ under the alternative hypothesis), making this selection procedure consistent contingent upon a suitable partition of the initial χ^2 values.

Lemma 2.1. *Under the alternative hypothesis there exists a subset A of* $[0,1]$ *with positive Lebesgue measure such that*

$$\left| \int_A \Delta(x)\,\mathrm{d}x - \int_{A^c} \Delta(x)\,\mathrm{d}x \right| > 0.$$

Proof. Under H_a, there exists a point (x_0, y_0) in $[0,1]^2$ and an $\varepsilon > 0$ such that $f(x_0, y_0) \neq f(x_0, 0)$ and $f(x_0, y_0) \neq f(x, y_0)$ for all $x \in I = [x_0 - \varepsilon, x_0 + \varepsilon]$. By the continuity of f, it is clear that $\Delta(x) > 0$ in an interval around x_0 which we call I_1. Consider $I^* = I \cap I_1$. It is clear that I^* is of the form $[s,t]$. Now, pick an interval A of the form $[s, s_1]$ and let $A^c = [s_1, t]$. If this selection gives $\int_A \Delta(x)\,\mathrm{d}x - \int_{A^c} \Delta(x)\,\mathrm{d}x = 0$, then choose a new A such that $A = [s, s_2]$, where $s_2 = (s_1 + t)/2$. This completes the proof of the lemma. □

In a similar manner, we can extend this procedure to the second stage, where we investigate the effect of a third predictor x_3, which is also assumed to take values in $[0,1]$, on the mean regression function. In this case, suppose the observations are of the form $\{Y_{ijk}, \mathbf{x}_{ij}^1, x_{3ijk}\}$, $k = 1, \ldots, n_{ij}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, N$. Here $\mathbf{x}_{ij}^1 = (x_{1i}, x_{2ij})'$ and it is assumed that the x's are ordered at each level $i \to j \to k$. Now, we form initial variance estimators χ_{ij}^2 as

$$\chi_{ij}^2 = \frac{1}{2(n_{ij} - 1)} \sum_{k=1}^{n_{ij}} [Y_{ijk} - Y_{ij1}]^2.$$

From these initial values, we form a statistic

$$T_1 = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{\sqrt{m_i}} \left[\sum_{j=1}^{m_i} \chi_{ij}^2 - \chi_{ij+m_i}^2 \right], \tag{4}$$

where $m_i = [n_i/2]$. It is easy to show that the asymptotic distribution of T_1 is also normal with mean 0 and a variance τ^2 under the assumption that $n_{ij} \to \infty$ whenever $n_i \to \infty$ and $N \to \infty$.

The extension of the algorithm to higher dimensions can be done in a similar manner. If the data is of the form $\{Y_{i_1 i_2 \ldots i_{j+1}}, \mathbf{x}_{i_1 i_2 \ldots i_j}^j, x_{i_{j+1}}\}$, where $j \leqslant p-1$, all the covariates are between 0 and 1 and, $\mathbf{x}_{i_1 i_2 \ldots i_j}^j = (x_{i_1}, \ldots, x_{i_j})'$, we can form the χ^2 statistics using the responses $Y_{i_1 i_2 \ldots i_{j+1}}$ at each $\mathbf{x}_{i_1 i_2 \ldots i_j}^j$ and group them in a suitable manner to construct the test statistic T_j. The asymptotic properties of the statistic remain the same as in Theorem 2.1 provided that the number of design points at each level tend to ∞.

Remark 2.3. The selection of the best two groups of the χ^2 values in higher dimensions is a critical issue. In a selection where the first-half of the χ^2 values in the jth direction are matched with the second-half, the parameter that detects any departures from the null hypothesis of no effect from the $(j+1)$st variable is

$$\int_{j-1} \left[\int_0^{1/2} \Delta(x_1, \ldots, x_j)\,\mathrm{d}x_j - \int_{1/2}^{1} \Delta(x_1, \ldots, x_j)\,\mathrm{d}x_j \right] \mathrm{d}x_1 \ldots \mathrm{d}x_{j-1},$$

where $\Delta(x_1, \ldots, x_j) = \int_0^1 [f(x_1, \ldots, x_j, x_{j+1}) - f(x_1, \ldots, x_j, 0)]^2\,\mathrm{d}x_{j+1}$ and $\int_j$ indicates the integral over $[0,1]^j$. The problem with trivial power can be much more serious when j gets large since it is difficult to judge

the behavior of the regression function in a higher dimension. This may lead to tests that are somewhat less powerful in a higher dimension (>3). Since $\Delta(x_1,\ldots,x_j)>0$ for all $\mathbf{x}$ in $[0,1]^j$, considering the volume under the hyper surface Δ, one can setup a region A in the jth direction as in case of Lemma 2.1 to obtain

$$\tilde{\Delta}(x_1,\ldots,x_{j-1})=\left|\int_A \Delta(x_1,\ldots,x_j)\,\mathrm{d}x_j-\int_{A^c}\Delta(x_1,\ldots,x_j)\,\mathrm{d}x_j\right|>0$$

so that $|\int_{j-1}\tilde{\Delta}(x_1,\ldots,x_{j-1})\,\mathrm{d}x_1\ldots\mathrm{d}x_{j-1}|>0$. However, locating the most appropriate region A would be somewhat difficult in higher-dimensional cases.

We can implement the above algorithm in a more general data setting. In particular, when we are examining the case of two covariates, suppose that the requirement of $n_{ij}>1$ at each i is removed. In this situation we can slice the covariate domain in the following manner. Select the covariate values $0<x^*_{11}<x^*_{12}<\cdots<x^*_{1M}<1$ from the predictor x_1 already in the model such that $M<N$ and $\max_{1\leqslant i\leqslant M+1}\{x^*_{1i}-x^*_{1\ i-1}\}=c_N$ for some $M\to\infty$ and $c_N\to 0$ as $N\to\infty$. Here $x_{10}=0$, $x_{1\ M+1}=1$. Now, we form χ^2 statistics by taking the difference between the responses that correspond to x_1 values within $I_i=[(x^*_{1i}+x^*_{1\ i-1})/2,(x^*_{1i}+x^*_{1\ i+1})/2]$ and the response at x^*_{1i} for each $i=2,\ldots,M-1$. For $i=1$ (M), we take the responses in the interval $I_1=[0,(x^*_{11}+x^*_{12})/2]$ $(I_M=[(x^*_{1M}+x^*_{1\ M-1})/2,1])$ in constructing the variance estimator. Specifically, if Y_{i^*1} is the response corresponding to x^*_{1i} and Y_{i^*j}, $j=2,\ldots,n_{i^*}$ are the other responses corresponding to the first covariate x_1 in the slice I_i, we set

$$\chi^2_i=\frac{\sum_{j=1}^{n_{i^*}-1}[Y_{i^*j}-Y_{i^*1}]^2}{2(n_{i^*}-1)},\quad i=1,\ldots,M.$$

Now we can form the test statistic to obtain

$$T=\frac{1}{\sqrt{M_1}}\sum_{i=1}^{M_1}\chi^2_i-\chi^2_{i+M_1},$$

where $M_1=[M/2]$.

Under an assumption that f has a continuous first derivative on $[0,1]^2$, we can establish the asymptotic normality of the test statistic as in Theorem 2.1 when the number of slices M increases with N. For higher dimensions, we proceed to define slices around a suitable collection of already selected variable combinations and define the test statistic in a similar manner. When we consider testing a hypothesis of type

$$\mathrm{H}_0\colon\ f(\mathbf{x}^j)=f_0(\mathbf{x}^{j-1}),$$

where $\mathbf{x}^j=(x_1,\ldots,x_j)'$, we can form the χ^2 statistics on a number of disjoint slices defined on $\mathbf{x}^{j-1}$. One way to define these slices is simply to examine $j-1$ dimensional squares of length $c_{n^{j-1}}$, where $n^j=\sum_j n_{i_1\ldots i_j}$.

2.2. Method 2

The second selection method we propose also depends on variance estimators. For ease in explanations, we begin with the two covariate case as in the beginning of the section using the same notation. Here, we compare two types of variance estimators under the null hypothesis $\mathrm{H}_0\colon f(x_1,x_2)=f_0(x_1)$. The first estimator $\hat{\sigma}^2_1$ is formed using χ^2 values for the first m slices as

$$\hat{\sigma}^2_1=\frac{1}{m}\sum_{i=1}^{m}\chi^2_i.$$

The second set of estimators would be formed using the differences method as follows. Let

$$\chi^2_{1i}=\frac{\sum_{j=1}^{n_i-r}[\sum_{k=0}^{r}d_kY_{ij+k}]^2}{n_i-r},\quad i=m+1,\ldots,n,$$

where d_k, $k = 0, \ldots, r$ is one of the optimal difference sequences (Hall et al., 1990) for variance estimation in nonparametric regression. Now, let

$$\hat{\sigma}_2^2 = \frac{1}{m} \sum_{i=1}^{m} \chi_{1i}^2.$$

Our test statistic would be $S_1 = \sqrt{m}(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)$, where we reject the null hypothesis for large values of S_1. Under the null hypothesis, the asymptotic distribution is given in the following theorem. The proof of this theorem follows directly from the proof of the previous theorem as well as results on difference estimators of nonparametric regression (Hall et al., 1990).

Theorem 2.2. *Under the null hypothesis and assumptions $\mathscr{A}_1$–$\mathscr{A}_3$, the statistic S converges in distribution to a normal random variable with mean* 0 *and variance* $\tau_1^2 = \tau^2/2 + \kappa\sigma^4 + 2\sigma^4 \sum_{j=1}^{n_i - r} [\sum_{k=0}^{r} d_k d_{j+k}]^2$, $\kappa = E(\varepsilon^4)\sigma^{-4} - 3$ *and* τ *is given in the previous theorem.*

Under the alternative hypothesis that $f = f(x_1, x_2)$, $E(S_1)/\sqrt{N} \propto \int_0^{0.5} \Delta(x)\,\mathrm{d}x$, where Δ is defined above. This shows that the test to reject the null hypothesis H_0: $f(x_1, x_2) = f_0(x_1)$ for large values of S_1 is consistent.

The generalization of this method to cases with more than two covariates as well as to cases with designs without repeated measurements can be done in the same manner as in Method 1. For example, the test to determine whether x_3 is useful in the model when x_1 and x_2 are already present, we can partition the (x_1, x_2) space into two disjoint regions A and A^c, where A is the lower half of the (x_1, x_2) domain. Following the same notation as in Method 1, we can form estimators of the error variance

$$\chi_{ij}^2 = \frac{1}{2(n_{ij} - 1)} \sum_{k=1}^{n_{ij}} [Y_{ijk} - Y_{ij1}]^2, \quad j = 1, \ldots, [n_i/2]$$

for slices in A. Then, using slices in A^c, another set of estimators of σ^2 of the variance is formed using the difference method statistics

$$\chi_{1ij}^2 = \frac{\sum_{l=1}^{n_{ij} - r} [\sum_{k=0}^{r} d_k Y_{ijl+k}]^2}{n_{ij} - r}, \quad j = m_i, \ldots, n_i,$$

where $m_i = [n_i/2]$. These can be combined as

$$S_2 = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{\sqrt{m_i}} \left[\sum_{j=1}^{m_i} \chi_{ij}^2 - \chi_{1ij}^2 \right]. \tag{5}$$

The null hypothesis H_0: $f(x_1, x_2, x_3) = f_0(x_1, x_2)$ would be rejected for large values of S_2. Under the same set of assumptions, following the same arguments, it can be shown that the asymptotic distribution of the test statistic under the null hypothesis in each case remain unchanged.

3. Empirical results

In this section, we present the results of a small simulation study and an example using some real data. We do not compare the selection methods proposed here with the methods proposed by Zhang (1991) and Härdle and Korostelev (1996) because these two methods are ad hoc methods that do not depend on a formal test and, as we mentioned earlier, sampling properties of these procedures under a null hypothesis of the type discussed here are unknown.

Table 1
Empirical power (times 1000) of size 0.05 tests for selection of x_2 when x_1 is in the model with normal errors

Function $f(x_1, x_2)$	N	n_i	$P1$	$P2$
0	10	10	75	22
	20	20	65	25
x_1^2	10	10	66	35
	20	20	63	24
$x_1 + x_2$	10	10	401	493
	20	20	415	703
$x_2 \sin(4x_1)$	10	10	314	261
	20	20	327	421
$x_1 x_2$	10	10	354	104
	20	20	439	194
$x_1 + \sin(4x_2)$	10	10	287	447
	20	20	328	831
$\cos(4x_1) + \sin(4x_2)$	10	10	319	429
	20	20	335	850
$\cos(4x_1)\sin(4x_2)$	10	10	282	174
	20	20	324	389

3.1. Simulations

We conducted a small simulation study to examine the power and size properties of the proposed selection methods. We examined the selection in two- and three-dimensional covariate settings using various types of regression functions. For simplicity, we examined the situations where multiple observations were available for each covariate combination that is, in two-dimensional data, we considered samples of type $\{Y_{ij}, X_{1i}, X_{2ij}\}$, $j=1,\ldots,n_i$, $i=1,\ldots,N$ and for three-dimensional covariates, we assumed the samples were of type $\{Y_{ijk}, \mathbf{x}_{ij}^1, x_{3ijk}\}$, $k=1,\ldots,n_{ij}$, $j=1,\ldots,n_i$, $i=1,\ldots,N$. The values of n_i, $i=1,\ldots,N$ and n_{ij} $j=1,\ldots,n_i$ in each setting were taken equal. Depending on the dimension, we examined several combinations of N, n_i and n_{ij} to get reasonably large total sample sizes. The error structure was taken to be normal with zero mean and a standard deviation 0.1. In most cases, we took equidistant covariate values. A few cases where the covariate values were generated using a uniform distribution on $[0, 1]$ were also studied. In the second method, we used $d_0 = -d_1 = 1/\sqrt{2}$ in variance estimation. All computations were done using S plus.

Tables 1 and 2 give a summary of the results of the simulations. In each table, $P1$ stands for the simulated power of Method 1 and $P2$ stands for the simulated power of the second method. It is clear that when the number of observations is large in the direction of the variable that is being examined, the power of the tests become large. Also, the power corresponding to a test for the effect of a second covariate seem to exceed the power of a test for the third covariate with the same type of contribution to the regression function by the additional covariate, which is not surprising. The results given here are for the case where the covariates are equidistant in all dimensions. The few cases where we used randomly generated covariates at each level did not produce results that were significantly different from the results we report here.

Table 2
Empirical power (times 1000) of size 0.05 tests for selection of x_3 when x_1 and x_2 are in the model with normal errors

Function $f(x_1,x_2,x_3)$	N	n_{ij}	n_{ijk}	$P1$	$P2$
0	5	5	10	64	18
	10	10	10	55	11
x_1+x_2	5	5	10	112	82
	10	10	10	68	5
$x_1+x_2+x_3$	5	5	10	322	551
	10	10	10	272	950
$x_1x_2x_3$	5	5	10	120	119
	10	10	10	215	120
$\sin(4(x_1+x_2+x_3))$	5	5	10	512	884
	10	10	10	545	1000
$[x_1^2+x_2^2+x_3^2]^{1/2}$	5	5	10	196	199
	10	10	10	197	202

Method 1 appears to have size little larger than the nominal level for sample sizes that were not very large. However, the difference shrinks as the sample size increases. Also, it seems that whether the null hypothesis is of the form H_0: $f(\mathbf{x})=0$ or it is of the form that the f is not a function of the variable being tested, the size does not change much. It is evident that Method 1 has lower power whenever the model is additive. This is not surprising. Method 2 seems to perform very well in most situations.

3.2. US temperature data example

We applied our procedures to US temperature data given in Hand et al. (1994). The data set gives the average January temperature for 56 US cities against the latitude (x_1) and longitude (x_2). It was shown in Peixoto (1990) that both variables play an important role in the regression of the temperatures against the covariates. A perspective plot of the smoothed temperatures for rescaled covariates is given in Fig. 1. The effect of longitude is evident from this plot also. To apply the procedure, we examined the residuals from the two-dimensional smoothing above and found no strong evidence against normality. We proceeded to estimate the error variance using a difference estimator (Hall et al., 1990) and used the fact $E(\varepsilon^4)=3\sigma^4$ to get an estimate of the asymptotic variance of the test statistics for testing the effect of longitude in the model in the presence of latitude. Because of assumed normality, both statistics will have the same null asymptotic variance. We divided the covariate latitude into eight equal groups after ordering the data according to the values of the latitude and calculated the test statistics for two groupings: the first 50% versus the last 50% and the middle 50% versus the end 50%. The test statistics in the two groupings for the Method 1 were 0.695(0.4902) and 1.995(0.0466), respectively, with associated p values in parentheses. In applying Method 2, with grouping the first 50% versus the last 50% gives a test statistic(p value) 1.69(0.0455) while grouping the middle 50% against the end 50% gives a test statistic(p value) of 2.34(0.0096). It is seen that the regression surface is near symmetric in both covariates and therefore the first grouping does not detect the significance of the covariate x_2 using Method 1. However, the second grouping detects the importance of the covariate x_2. As expected, Method 2 is not affected by the symmetry. This example shows the importance of proper grouping in identifying important covariates in a model.

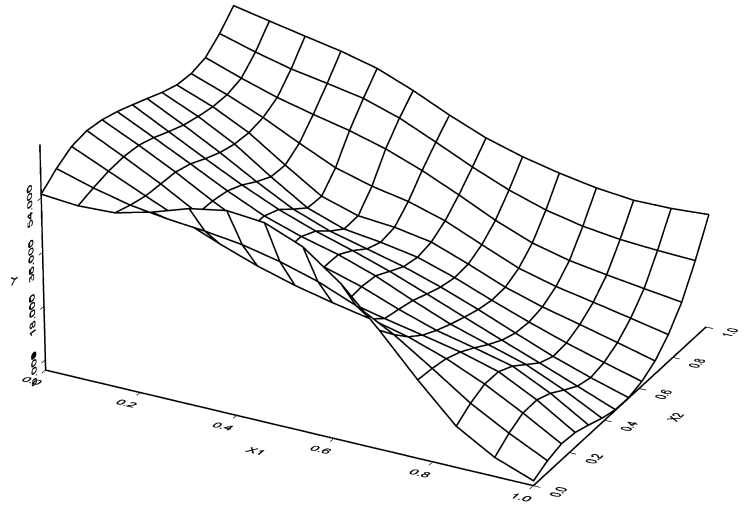

Fig. 1. Smoothed US temperature data.

The ineffectiveness of the Method 1 with the first grouping may also have an effect due to small number of slices as well as the small sample size within each slice (seven responses in each slice). Also, we feel that using the square of the error variance estimate to estimate the fourth moment of the error distribution produces an inflated asymptotic variance for the test statistics. For a few samples of experimental bandwidths we examined the moment estimators of $E(\varepsilon^4)$ based on the residuals from two-dimensional fits and most of them seem to be smaller than the value we found by squaring the estimator of the error variance obtained using the difference method.

4. Conclusion

The procedures discussed above rely solely on the fact that the regression curve has some change in the direction of a particular covariate. The grouping with respect to the variables that are already in the model plays a very crucial role in detecting the importance of a new variable in the model. We did not examine the best grouping for a given class of regression functions in this paper.

An alternative method of using the above selection method is to use single index models (Härdle et al., 1993) at each stage of the selection process so that one always has two variables to examine, one, a linear combination of the predictors that are already in the model and the new variable. The visual examination that may enable one to choose the best grouping may be worth that extra effort.

It may also be possible for one to examine the jth variable only with respect to the $(j-1)$st variable. For example, in examining the effect of the third covariate x_3 in the presence of x_1 and x_2, one may just use

$$T_1(x_1) = \frac{1}{\sqrt{m_i}} \left[\sum_{j=1}^{m_i} \chi_{ij}^2 - \chi_{ij+m_i}^2 \right],$$

as a test statistic for a fix value x_1 as opposed to (4). This could be done after adjusting for x_1. One disadvantage of such an approach is that the function $\Delta(x_1, x_2)$ may be very close to zero for some value of x_1 depending on the underlying function f and it would be difficult to check such behavior of the regression function.

Acknowledgements

This research is partially supported by a grant from Office of Naval Research and a grant from the NIH. The author wishes to thank Dr. A. Manatunga at Emory University and Dr. K. Selvavel at the Census Bureau for informative discussions during the preparation of the article.

References

Friedman, J.H., Stuetzle, W., 1981. Projection pursuit regression. J. Amer. Statist. Assoc. 760, 817–823.

Hall, P., Kay, J.W., Titterington, D.M., 1990. Asymptotically optimal difference based estimation of variance in nonparametric regression. Biometrika 77, 521–528.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., Ostrowski, E., 1994. Small Data Sets. Chapman & Hall, London.

Härdle, W., Hall, P., Ichimura, H., 1993. Optimal smoothing in single-index models. Ann. Statist. 21, 157–178.

Härdle, W., Korostelev, A., 1996. Search for significant variables in nonparametric additive regression. Biometrika 83, 541–550.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London.

Kulasekera, K.B., Wang, J., 1997. Smoothing parameter selection for power optimality in testing of regression curves. J. Amer. Statist. Assoc. 92, 500–511.

Li, K.C., 1991. Sliced inverse regression for dimension reduction. J. Amer. Statist. Assoc. 86, 316–327.

Peixoto, J.L., 1990. A Property of well formulated polynomial regression models. Amer. Statist. 44, 26–30.

Zhang, P., 1991. Variable selection in nonparametric regression with continuous covariates. Ann. Statist. 19, 1869–1882.