International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:11, No:3, 2017

# A Comparative Study of Additive and Nonparametric Regression Estimators and Variable Selection Procedures

Adriano Z. Zambom, Preethi Ravikumar

***Abstract*—**One of the biggest challenges in nonparametric regression is the curse of dimensionality. Additive models are known to overcome this problem by estimating only the individual additive effects of each covariate. However, if the model is misspecified, the accuracy of the estimator compared to the fully nonparametric one is unknown. In this work the efficiency of completely nonparametric regression estimators such as the Loess is compared to the estimators that assume additivity in several situations, including additive and non-additive regression scenarios. The comparison is done by computing the oracle mean square error of the estimators with regards to the true nonparametric regression function. Then, a backward elimination selection procedure based on the Akaike Information Criteria is proposed, which is computed from either the additive or the nonparametric model. Simulations show that if the additive model is misspecified, the percentage of time it fails to select important variables can be higher than that of the fully nonparametric approach. A dimension reduction step is included when nonparametric estimator cannot be computed due to the curse of dimensionality. Finally, the Boston housing dataset is analyzed using the proposed backward elimination procedure and the selected variables are identified.

***Keywords*—**Additive models, local polynomial regression, residuals, mean square error, variable selection.

## I. Introduction

IN nonparametric regression analysis, the focus is on studying and exploring the relationship between a possible set of independent variables denoted by $\mathbf{X} = (X_1, ..., X_d)$ and the response variable $Y$ through a regression function $m(\cdot)$ [1]. In the simple case, $\mathbf{X}$ can be considered as a fixed design with observation in a grid, but in the general case we consider $\mathbf{X}$ to be random with a multivariate density function $f_{\mathbf{X}}(\mathbf{x})$ (see [2], [3]).

The function $m$ is called the regression function, usually considered to be a smooth function such that

$$m(\mathbf{X}) = E(Y|\mathbf{X}). \tag{1}$$

It is typically assumed that the conditional variance of $Y$ given $\mathbf{X}$ is constant, but in a more general case, we can relax this assumption and let the variance depend on the explanatory variables $\mathbf{X}$. Thus, the general heterocedastic nonparametric regression model is [4]

$$Y_i = m(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\epsilon_i,\ i = 1...n \tag{2}$$

where $\epsilon_i$ are independent identically distributed random variables with mean 0 and constant variance.

A. Z. Zambom is with the Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL, 60660 USA (e-mail: azambom@luc.edu).
P. Ravikumar is with Loyola University Chicago.

Exploring relationships between variables is widely used when applied statistics is used in many different areas of research. Regression tools are hence necessary for the analysis of data coming from experiments in biology, engineering, chemistry, physics, and others. Constantly, we search for techniques of generalizing the models so that they can be useful in more complicated scenarios, and thus the goal is to relax assumptions to obtain less restrictive models while maintaining flexibility to incorporate a wide range of practical applications.

Recently a great amount of data is being collected due to the advance of computer technology. Various datasets containing variables with non-trivial/nonlinear relationships are more and more common, and hence the use of flexible models is crucial (see [5], [6]). The flexibility of a regression model is one of the fundamental aspects of modeling, together with the dimensionality and interpretability [7]. When focusing in flexibility, if the dimension of the variables is large, fitting the model becomes problematic, causing the well-known curse of dimensionality [8]. Moreover, not only should the model be flexible and deal with higher dimensions, but also have easy interpretation for the researcher. In this paper we focus mainly on the issue of flexibility, where model checking and variable selection can be performed without strong restrictions or assumptions on the regression function.

Nonparametric regression was initially developed as a way of generalizing the classical parametric regression

$$Y_i = m(\theta, \mathbf{X}) + \epsilon_i,\ i = 1...n \tag{3}$$

where $\beta$ are the parameters and $m(\theta, \mathbf{X})$ is of a parametric form. This model generally assumes that the family of models $\{m(\theta, \mathbf{x}), \theta \in \Theta\}$ contains the true underlying model that generated the data. Such restriction can cause serious disadvantages, yielding disastrous estimations and prediction if the assumptions do not hold [9].

The first and most commonly used parametric regression model, a particular case of (3), is the parametric linear regression [10], which assumes that the effects of the predictors on the response variable are additive and linear, which is a very strong assumption. Even though this is very restrictive, it is widely used and does apply to many situations, there is an increasing amount of data from several areas that can not be fit by a linear model, see [11]-[14] just to cite a few. An alternative of the linear model that considers nonlinear effects of the independent variables adds powers of the covariates as linear regressors, however the number of

parameters to estimate increases and leads to an over-fit and lack of interpretability when many powers are required.

In this paper we study the contrast between flexibility and interpretability of nonparametric models, more specifically, those who consider the fully nonlinear model on all predictors and the additive regression model. The idea is to compare these techniques when they are correctly and incorrectly specified.

## II. Nonparametric Methods

There are many alternatives for linear regression to consider nonlinear effects of the independent variables. The Nadaraya-Watson Kernel estimator [15], [16] was one of the first methods to establish this technique. It is actually a special case of the Local Polynomial regression [17]-[21]. Consider first the univariate case where $X$ is uni-dimensional. In estimating of $m(x)$ with Local Polynomial Regression, the classical weighted least squares regression is used to fit a $q$ degree polynomial

$$\beta_0 + \beta_1(.-x) + ... + \beta_p(.-x)^q$$

to the data $(X_i, Y_i)$, where the weights are the kernel functions $K_{h_n}(X_i - x) = \frac{1}{h_n}K(\frac{X_i-x}{h_n})$, for a function $K$ such that $\int K(x)dx = 1$ called the kernel function (usually taken to be a density). Therefore, the goal is to minimize

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1(X_i-x) - \ldots - \beta_p(X_i-x)^q\right)^2 K_{h_n}(X_i-x) \tag{4}$$

with respect to $(\beta_0, \ldots, \beta_p)$. The estimated mean regression function $m(\cdot)$ is computed by $\hat{\beta}_0$. In order to derive this estimator it is assumed that the function $m$ has $q+1$ derivatives continuous at $x$. This assumption is necessary because a Taylor expansion $m(X_i) \approx m(x) + m'(x)(X_i - x) + \ldots + \frac{m^{(q)}(x)}{q!}(X_i-x)^q$ is used as an approximation to $m(X_i)$. For simplicity, assume that the support of $f_X$ is on [0,1] and the kernel is supported on [-1,1]. Also assume that $m''$, $f'$, and $\sigma$ are continuous, the kernel is symmetric about 0 and the bandwidth $h_n$ is such that $h_n \to 0$ and $nh_n \to \infty$. In matrix form, let $Y = (Y_1, \ldots, Y_n)'$,

$$X_x = \begin{pmatrix} 1 & X_1 - x & \ldots & (X_1-x)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \ldots & (X_n-x)^q \end{pmatrix} \tag{5}$$

and $W_x = \text{diag}\{K_{h_n}(X_1 - x), \ldots, K_{h_n}(X_n - x)\}$ is the diagonal matrix of weights. It is clear that the solution of this weighted least squares problem is

$$\hat{\beta} = (X_x^T W_x X_x)^{-1} X_x^T W_x Y \tag{6}$$

assuming that $(X_x' W_x X_x)$ is invertible. The estimator of $m(x)$ is

$$\hat{m}(x; q, h_n) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x Y, \tag{7}$$

where $e_1 = (1, 0, \ldots, 0)$ of size $(q+1) \times 1$.

Now consider the case where $\mathbf{X}$ is $d$-dimensional. Assume that the regression function $m(\mathbf{z})$ is such that $q+1$ derivatives exist and are continuous at $\mathbf{x}$. Then we can approximate $m(\mathbf{z})$ by a multivariate polynomial

$$m(\mathbf{z}) \approx \sum_{0 \le |\mathbf{k}| \le q} \frac{1}{\mathbf{k}!} D^{\mathbf{k}} m(\mathbf{y})|_{\mathbf{y}=\mathbf{x}} (\mathbf{z}-\mathbf{x})^{\mathbf{k}} \tag{8}$$

where

$$\mathbf{k} = (k_1, \ldots, k_d), \quad \mathbf{k}! = k_1! \times \ldots \times k_d!, \quad |\mathbf{k}| = \sum_{i=1}^{d} k_i$$

$$\mathbf{x}^{\mathbf{k}} = (x_1^{k_1} \times \ldots \times x_d^{k_d}), \sum_{0 \le |\mathbf{k}| \le q} = \sum_{j=0}^{q} \sum_{\substack{k_1=0 \\ k_1+\ldots+k_d=j}}^{j} \cdots \sum_{k_d=0}^{j},$$

$$(D^{\mathbf{k}} m)(\mathbf{y}) = \frac{\partial^{\mathbf{k}} m(\mathbf{y})}{\partial y_1^{k_1} \ldots \partial y_d^{k_d}}.$$

In this case, the solution is the minimizer of the multivariate least squares

$$\sum_{i=1}^{n}\left(Y_i - \boldsymbol{\beta}'(\mathbf{X}_i - \mathbf{x})^{\mathbf{k}}\right)^2 K_{H_n}(\mathbf{X}_i - \mathbf{x}), \tag{9}$$

where $H_n$ is the $d \times d$ bandwidth matrix, assumed to be symmetric and positive definite and the kernel $K_{H_n}$ is a $d$-variate form of the kernel $K$. The kernel can have many forms, including product of univariate kernels or $d$-variate probability density functions, but usual assumptions are that $\int K(\mathbf{u})d\mathbf{u} = 1$ and $K_{H_n}(\mathbf{u}) = |H_n|^{-1/2}K(H_n^{-1/2}\mathbf{u})$. See [22] for details.

In the following example, we will demonstrate the performance of nonparametric regression models and the parametric linear regression model.

**Example 1:** Suppose that we have pairs of observations $(Y_i, X_i)$ from the unknown models Model 1: $Y = 6X\cos(6\pi X) + \epsilon$ and Model 2: $Y = 4X\cos(5\pi X) + \epsilon$, where $X \sim U(0,1)$ and $\epsilon \sim N(0,1)$. Table I shows the residual sums of squares for the simple linear regression (SLR) including the powers of the independent variable up to $p$, local linear and Splines [1]. The results reported in Table I suggest that the nonparametric models (local linear and splines) yield much lower residual sum of squares compared to the parametric simple linear regression. Even with the increase of the polynomial order of the variables used, the simple linear regression continues to achieve higher residual sum of squares compared to the nonparametric methods. This demonstrates the great importance of having a flexible model that does not restrict the estimation to a small range of parametric families, when the true model can have nonlinear effects on the response. A widely used alternative to the local regression fitting is a generalization of the parametric

TABLE I
Comparison of RSS for Different Methods of Regression Estimation

| | SLR (p=1) | SLR (p=2) | SLR (p=5) | Local Linear | Splines |
|---|---|---|---|---|---|
| Model 1 | 737.1 | 703.3 | 543.3 | 83.1 | 69.5 |
| Model 2 | 414.4 | 410.4 | 210.5 | 89.9 | 61.6 |

linear regression. [23], [24] considered this case, known as nonparametric additive models, where

$$Y = m_1(X_1) + m_2(X_2) + ... + m_d(X_d) + \epsilon, \qquad (10)$$

for the covariates $X_1, ..., X_d$ (see also [25]). For identifiability of the model, it is required that $E(X_i) = 0, i = 1, \ldots, d$. The effect of the covariates on the response variable in this model is not linear, but it is additive. This relates each explanatory variable $X_j$ to the response $Y$ in an additive way, but the functions $m_i$ are not parametrically specified in advance, but is determined by the data analytically through local smoothing. Furthermore, one fundamental property of the linear model is retained: easy interpretation. The advantage here is that, after fixing the values of all other covariates, the effect of a specific covariate on the response is a curve that can be seen as univariate.

Once again the trade off between interpretability and flexibility comes into play, especially if the number of available covariates is high. The additive model maintains the asymptotic convergence rates of a single predictor, no matter how many dimensions (number of predictors). However it does not incorporate more complex models, as for example interactions, ratios, or any non-additive function of two or more covariates. If an interaction is in the true underlying model for instance, the change in the response variable is more complex than just an additive effect of variables. In fact, in practical applications found in several scientific areas interaction terms are commonly identified and are in fact critical to the inferential conclusions.

We can find many cases in applied statistics where the additive model may not be the correct one. For example, if we have a study of HDL Cholesterol predicted by BMI (body mass index) and Total Cholesterol, the additive effect of the predictors may not be correct, since BMI probably has a interaction with Total Cholesterol. Another example would be the regression of two different medications in a situation where covariates are the levels of different treatments (in medicine or agriculture) and the response is some measure of a reaction against a particular disease.

Consider the additive model in (10), where we assume without loss of generality that the $Y_i$ has expected value 0. Note that in this model, for any $k = 1 \ldots d$,

$$E(Y - \sum_{j \neq k}^{d} m_j(X_j)|X_k) = m_k(X_k). \qquad (11)$$

Following [26], let $\mathbf{m}_j = (m_d(X_{j1}), \ldots, m_j(X_{jn}))^T$ be the vector corresponding to the function $m(\cdot)$ at the observed covariate values. These additive components are estimated by solving the set of normal equations

$$\begin{bmatrix} I & \mathbf{S}_1 & \ldots & \mathbf{S}_1 \\ \mathbf{S}_2 & I & \ldots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_d & \mathbf{S}_d & \ldots & I \end{bmatrix} \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_d \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_d \end{bmatrix} Y, \qquad (12)$$

where $\mathbf{S}_j$ is the linear smoother matrix with respect to the $\mathbf{X}_j = (X_{j1}, \ldots, X_{jn})$ covariate vector.

Intuitively from this formulation, an iterative algorithm is designed in order to solve these equations. This well-know algorithm is called the backfitting algorithm, which has been shown to converge to the solution

$$\begin{bmatrix} \hat{\mathbf{m}}_1 \\ \hat{\mathbf{m}}_2 \\ \vdots \\ \hat{\mathbf{m}}_d \end{bmatrix} = \begin{bmatrix} I & \mathbf{S}_1 & \ldots & \mathbf{S}_1 \\ \mathbf{S}_2 & I & \ldots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_d & \mathbf{S}_d & \ldots & I \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_d \end{bmatrix} Y = M^{-1}CY, \qquad (13)$$

if $M^{-1}$ exists. Hence, the backfitting algorithm can be described in the following way

1) Initialize: $\hat{\mu} = \bar{Y}$, set $m_j = 0$, $j = 1 \ldots p$
2) for $j = 1 \ldots d$
   $\hat{m}_k = S_k(Y - \hat{\mu} - \sum_{j \neq k}^{d} m_j | X_k)$
3) repeat 2) until convergence

In each step of the algorithm, the estimated mean function $m_j$ is readjusted, after the removal of the effects of the other predictors. This takes advantage of the additive effects with the partial residuals in each iteration. This algorithm is similar to the well-known Gauss-Seidel algorithm, which iterates the algorithm in the same way, readjusting each step. Here we set the initial functions $m_j$ to be equal to 0, but a better idea would be to start these functions with some previous knowledge or even a simple fit (simple linear regression for instance).

Using the notation in [26], in order to have an interpretable expression for each $\hat{m}_j$, define the additive smoother matrix $W_j = E_j M^{-1} C$, where $E_j$ is a partitioned $n \times nd$ matrix with the j-th block being a $nxn$ identity matrix and zeros elsewhere. In this way, we have $\hat{m}_j = W_j Y$ and $\hat{m} = \sum_{j=1}^{d} \hat{m}_j = WY$, for $W = W_1 + \ldots + W_d$.

Other models can be derived as variations of the additive model, and those include semi-parametric models, varying-coefficient models, partially linear models, varying-coefficient partially linear models, etc. In all of these cases, the effect of each covariate is assumed to be additive, and a more general model is always of interest.

## III. Numerical Results

In this section, we analyze the finite sample performance of the additive models and the local polynomial estimators (degree of polynomial 1) in several scenarios. The additive model was computed using the *mgcv* package and the local polynomial using the *loess* function in the statistical software R. The goal is to study their results under several models, some whose underlying generating process is additive and other with interactions. All data in this section is generated from the general nonlinear model

$$Y = m_k(\mathbf{X}) + \epsilon, k = 1, \ldots, 10,$$

where $\epsilon$ is Normally distributed with mean value 0 and variance $\sigma^2 = 1$. The independent variables $\mathbf{X}$ are generated from a Uniform distribution with support [0,5]. First, a sample of size $n_{train}$ is generated and the models are constructed. The

first measure of comparison for the additive and local linear methods we consider is the residual sum of squares

$$RSS = \sum_{i=1}^{n_{train}} (Y_i - \hat{m}_k(X_i))^2,$$

where $\hat{m}(\cdot)$ is the regression estimator, either additive using the backfitting algorithm or local polynomial in all dimensions. Note that the comparison of RSS is standard in regression analysis, however it is only fair if the selection of the tuning parameters, i.e. bandwidths, are properly chosen for comparison. If the bandwidths are chosen too small, then the RSS can be as close to 0 as desired, induced by interpolation. This is the case of over-fitting, causing poor prediction of new observations. In order to assess the ability of the models to predict new observations, we also report what we call the oracle RSS. After constructing the model with $n_{train}$, we generate $n_{test}$ samples $\mathbf{X}^{new}$ from the original model and compare the predictions at the new observations with the actual values. This measure is defined as

$$ORSS = \sum_{i=1}^{n_t} (\hat{m}_k(\mathbf{X}_i^{new}) - m_k(\mathbf{X}_i^{new}))^2.$$

Obviously in a real data situation the original function $m_k(\cdot)$ is not known. However in this simulation setting we use it as an oracle way of comparing the performance of the two methods. We exclude from the results those $\mathbf{X}^{new}$ that are outside the range of the original ones since the models do not predict outside the range.

In this simulation section, we study two scenarios: The first with 2 independent covariates and the second with 3 independent covariates. In both scenarios we generate additive and non-additive models in order to analyze how much gain/loss is obtained by using the correct/wrong model in each case. The models for two-dimensional scenario are:

$$\begin{aligned}
m_1(X_1, X_2) &= X_1 + X_2, \\
m_2(X_1, X_2) &= X_1^2 + X_2, \\
m_3(X_1, X_2) &= X_1^3 + X_2^2, \\
m_4(X_1, X_2) &= \sin(X_1) + X_2, \\
m_5(X_1, X_2) &= X_1 X_2, \\
m_6(X_1, X_2) &= X_1 \sin(X_2), \\
m_7(X_1, X_2) &= \sin(X_1)\sin(X_2), \\
m_8(X_1, X_2) &= X_1 / X_2, \\
m_9(X_1, X_2) &= X_1^3 X_2, \\
m_{10}(X_1, X_2) &= \exp(X_1 - X_2).
\end{aligned}$$

The average results of 1000 simulation runs for the scenario with two covariates can be found in Table II. For the additive models $m_1, \ldots, m_4$, the residual sum of squares of both methods are very similar, however the additive regression method obtains smaller oracle residual sum of squares in all cases. On the other hand, the local polynomial estimator obtains much smaller RSS and ORSS in all non-additive models $m_5, \ldots, m_{10}$. In fact for small sample size 50, the ORSS achieved by the local polynomial estimator is 3.55 times smaller than that obtained by the additive regression in average for $m_5, \ldots, m_{10}$, while only 1.28 times larger for $m_1, \ldots, m_4$. For large sample size 300 the differences are even larger for $m_5, \ldots, m_{10}$, where the local polynomial has in average 5 times smaller ORSS compared to that of the additive model, while yielding only 1.3 larger ORSS for $m_1, \ldots, m_4$. This suggests that the gain of using additive nonparametric regression for this simulated scenario is small when the true underlying model is additive, while the loss is much larger for non-additive cases.

TABLE II
MSE AND MSE ORACLE FOR MODELS WITH 2 COVARIATES

| $n_{train}$ | Model | Additive | | Local Polynomial | |
|---|---|---|---|---|---|
| | | RSS | ORSS | RSS | ORSS |
| 50 | $m_1$ | 40.9 | 13.5 | 37.5 | 16.2 |
| | $m_2$ | 40.1 | 14.4 | 37.7 | 16.5 |
| | $m_3$ | 75.7 | 30.2 | 78.4 | 47.0 |
| | $m_4$ | 39.9 | 14.2 | 38.4 | 17.4 |
| | $m_5$ | 196.1 | 72.8 | 37.3 | 16.1 |
| | $m_6$ | 80.8 | 40.0 | 42.3 | 21.6 |
| | $m_7$ | 49.6 | 24.1 | 38.5 | 18.1 |
| | $m_8$ | 213313.6 | 323.6 | 178711.9 | 232.0 |
| | $m_9$ | 96591.8 | 1685.6 | 1067.5 | 193.5 |
| | $m_{10}$ | 3046.5 | 246.6 | 283.7 | 70.4 |
| 150 | $m_1$ | 139.3 | 8.3 | 137.7 | 9.4 |
| | $m_2$ | 140.6 | 9.0 | 137.6 | 9.4 |
| | $m_3$ | 264.8 | 27.8 | 300.3 | 46.2 |
| | $m_4$ | 139.4 | 8.5 | 138.9 | 10.4 |
| | $m_5$ | 728.7 | 76.5 | 136.5 | 9.5 |
| | $m_6$ | 285.8 | 40.2 | 154.8 | 16.7 |
| | $m_7$ | 175.3 | 21.4 | 142.5 | 11.8 |
| | $m_8$ | 11854754 | 460.5 | 11119423 | 380.1 |
| | $m_9$ | 359558.2 | 1788.3 | 4255.1 | 194.5 |
| | $m_{10}$ | 10923.3 | 265.7 | 1197.5 | 71.5 |
| 300 | $m_1$ | 289.9 | 5.8 | 288.2 | 6.6 |
| | $m_2$ | 291.1 | 6.8 | 286.6 | 6.8 |
| | $m_3$ | 543.7 | 27.2 | 639.6 | 47.0 |
| | $m_4$ | 290.2 | 6.1 | 291.1 | 8.1 |
| | $m_5$ | 1530.7 | 77.4 | 286.8 | 6.8 |
| | $m_6$ | 599.5 | 40.4 | 328.9 | 15.3 |
| | $m_7$ | 362.8 | 20.9 | 295.8 | 9.8 |
| | $m_8$ | 156429178 | 815.3 | 141777570 | 857.2 |
| | $m_9$ | 746330.2 | 1830.3 | 9126.4 | 198.7 |
| | $m_{10}$ | 23088.1 | 266.7 | 2659.6 | 71.9 |

The models for three-dimensional case are:

$$\begin{aligned}
m_1(X_1, X_2) &= X_1 + X_2 + X_3, \\
m_2(X_1, X_2) &= \cos(X_1) + \exp(X_2) - \exp(X_3), \\
m_3(X_1, X_2) &= X_1^2 + X_2^3 + X_3, \\
m_4(X_1, X_2) &= \sin(X_1) + X_2 + \sin(X_3), \\
m_5(X_1, X_2) &= X_1 X_2 X_3, \\
m_6(X_1, X_2) &= \sin(X_1) + \sin(X_2)\sin(X_3), \\
m_7(X_1, X_2) &= X_1 + X_2^2 * X_3^{1/2}, \\
m_8(X_1, X_2) &= X_1 + X_2 * X_3, \\
m_9(X_1, X_2) &= \sin(X_1 + X_2 + X_3), \\
m_{10}(X_1, X_2) &= \exp(X_1/4 + X_2/4 - X_3/2).
\end{aligned}$$

The average results of 1000 simulation runs is reported on Table III. The same pattern observed in the two-dimensional cases is observed for the three-dimensional case. As the sample size increases the local polynomial estimator seems

to significantly improve (lower ORSS) for most additive and non-additive models, while the additive regression only improves for additive models. Note that, for sample size 300, the interaction model $m_5$ is extremely difficult for the additive regression. The ORSS is almost 18 times larger than that of the local polynomial regression.

TABLE III
MSE AND MSE ORACLE FOR MODELS WITH 3 COVARIATES

| $n_{train}$ | Model | Additive | | Local Polynomial | |
|---|---|---|---|---|---|
| | | RSS | ORSS | RSS | ORSS |
| 50 | $m_1$ | 35.2 | 16.2 | 27.3 | 24.4 |
| | $m_2$ | 598.9 | 118.2 | 1283.1 | 278.1 |
| | $m_3$ | 70.9 | 33.5 | 97.4 | 71.2 |
| | $m_4$ | 35.3 | 16.6 | 29.3 | 26.2 |
| | $m_5$ | 2957.6 | 322.7 | 34.9 | 36.0 |
| | $m_6$ | 44.2 | 25.2 | 29.4 | 27.3 |
| | $m_7$ | 516.8 | 130.4 | 35.5 | 32.5 |
| | $m_8$ | 178.9 | 74.1 | 27.2 | 24.2 |
| | $m_9$ | 51.6 | 32.8 | 32.9 | 32.6 |
| | $m_{10}$ | 44.8 | 24.4 | 27.8 | 24.7 |
| 150 | $m_1$ | 133.6 | 9.8 | 124.2 | 13.0 |
| | $m_2$ | 2105.3 | 115.4 | 6139.1 | 275.5 |
| | $m_3$ | 256.7 | 29.0 | 452.1 | 70.8 |
| | $m_4$ | 137.4 | 10.2 | 132.4 | 16.0 |
| | $m_5$ | 11952.0 | 335.5 | 154.1 | 22.4 |
| | $m_6$ | 168.5 | 22.5 | 133.1 | 17.2 |
| | $m_7$ | 2134.9 | 138.9 | 164.4 | 24.5 |
| | $m_8$ | 699.9 | 76.6 | 126.1 | 12.9 |
| | $m_9$ | 202.0 | 31.9 | 154.4 | 24.1 |
| | $m_{10}$ | 172.8 | 21.5 | 126.3 | 13.5 |
| 300 | $m_1$ | 282.0 | 7.2 | 273.3 | 9.0 |
| | $m_2$ | 4382.3 | 113.1 | 13721.0 | 278.8 |
| | $m_3$ | 535.4 | 28.4 | 1001.3 | 72.1 |
| | $m_4$ | 284.4 | 7.6 | 288.5 | 13.0 |
| | $m_5$ | 25869.1 | 338.2 | 333.4 | 19.1 |
| | $m_6$ | 354.4 | 21.4 | 296.3 | 14.8 |
| | $m_7$ | 4620.4 | 139.1 | 366.5 | 22.5 |
| | $m_8$ | 1507.8 | 77.3 | 275.4 | 8.9 |
| | $m_9$ | 423.2 | 31.6 | 337.4 | 22.6 |
| | $m_{10}$ | 372.1 | 20.2 | 275.3 | 9.7 |

## IV. VARIABLE SELECTION

Now suppose we have a set of available predictors $\mathbf{X} = (X_1, \ldots, X_d)$, however only a subset of them, whose indices we denote by $I_0 = (i_1, \ldots, i_{d_0})$ are significant in predicting the response $Y$.

In order to select the predictors, the classical linear regression uses stepwise procedures such as the backward elimination of the forward selection. These methods are usually based on the AIC, BIC or Mellow's Cp criteria to add or exclude a variable from the model at each step. The algorithm usually stops when adding or excluding a variable does not decrease the chosen criterium. We propose using a backward elimination procedure for the local polynomial regression and for the additive regression based on the AIC computed from these procedures [27], [24].

We simulated data from the model

$$Y = m(\mathbf{X}) + \epsilon,$$

where $\mathbf{X} = (X_1, \ldots, X_4)$, $X_j, j = 1, \ldots, 4$ are i.i.d. Uniform in [0,5], and $\epsilon \sim N(0, 1)$. The average of correctly selected models for 1000 simulation runs is shown in Table IV for several different models. Note that out of 4 total available predictors, only 2 or 3 of them should be selected as significant depending on the true underlying data generating model. Clearly, the local polynomial regression always selects all important predictors and never selects unwanted ones. On the other hand, the additive regression always selects significant predictors but in 15 to 42 percent of the time includes predictors that are not significant.

TABLE IV
MSE AND MSE ORACLE FOR MODELS WITH 3 COVARIATES

| | Model | Additive | | Local Polynomial | |
|---|---|---|---|---|---|
| | $m(\mathbf{X})$ | correct | incorrect | correct | incorrect |
| n = 100 | $X_1 + X_2$ | 2 | .42 | 2 | 0 |
| | $X_1 X_2$ | 2 | .38 | 2 | 0 |
| | $X_1 + X_2 + X_3$ | 3 | .17 | 3 | 0 |
| | $X_1 X_2 X_3$ | 3 | .20 | 3 | 0 |
| n = 200 | $X_1 + X_2$ | 2 | .32 | 2 | 0 |
| | $X_1 X_2$ | 2 | .39 | 2 | 0 |
| | $X_1 + X_2 + X_3$ | 3 | .19 | 3 | 0 |
| | $X_1 X_2 X_3$ | 3 | .15 | 3 | 0 |

## V. REAL DATA ANALYSIS

We applied the variable selection with backward elimination and local polynomial fit to the Boston housing dataset [28]. The variable $Y$ = median value of owner-occupied homes in \$1000's was used and the predictors were: Per capita crime rate by town, proportion of residential land zoned for lots over 25,000 sq.ft., proportion of non-retail business acres per town, nitric oxides concentration (parts per 10 million), average number of rooms per dwelling, proportion of owner-occupied units built prior to 1940, weighted distances to five Boston employment centres, pupil-teacher ratio by town, 1000(Bk - 0.63)$^2$ where Bk is the proportion of blacks by town, and % lower status of the population.

The proposed procedure selected: Proportion of residential land zoned for lots over 25,000 sq.ft., proportion of non-retail business acres per town, average number of rooms per dwelling, proportion of owner-occupied units built prior to 1940, weighted distances to five Boston employment centres, pupil-teacher ratio by town, 1000(Bk - 0.63)$^2$ where Bk is the proportion of blacks by town, and % lower status of the population, that is, only two predictors were excluded. An interesting graph is the plot of the weighted distances to five Boston employment centres and the response Median value of owner-occupied homes in \$1000's shown in Fig. 1. The non-linear relationship of these two variable is very easily seen, which makes nonparametric regression an essential tool in this analysis.

## VI. CONCLUSION

In the search, for regression models that demonstrate easy interpretation of the relationship of the predictors with the response, the challenge of avoiding misspecification is of great importance. In this paper the loss in predicting future observations when using the additive model was investigated in comparison with a full non-parametric local polynomial regression. Although a simpler model and with additive effects of each predictor, the additive regression obtained much larger residual sum of squares of future observations for all

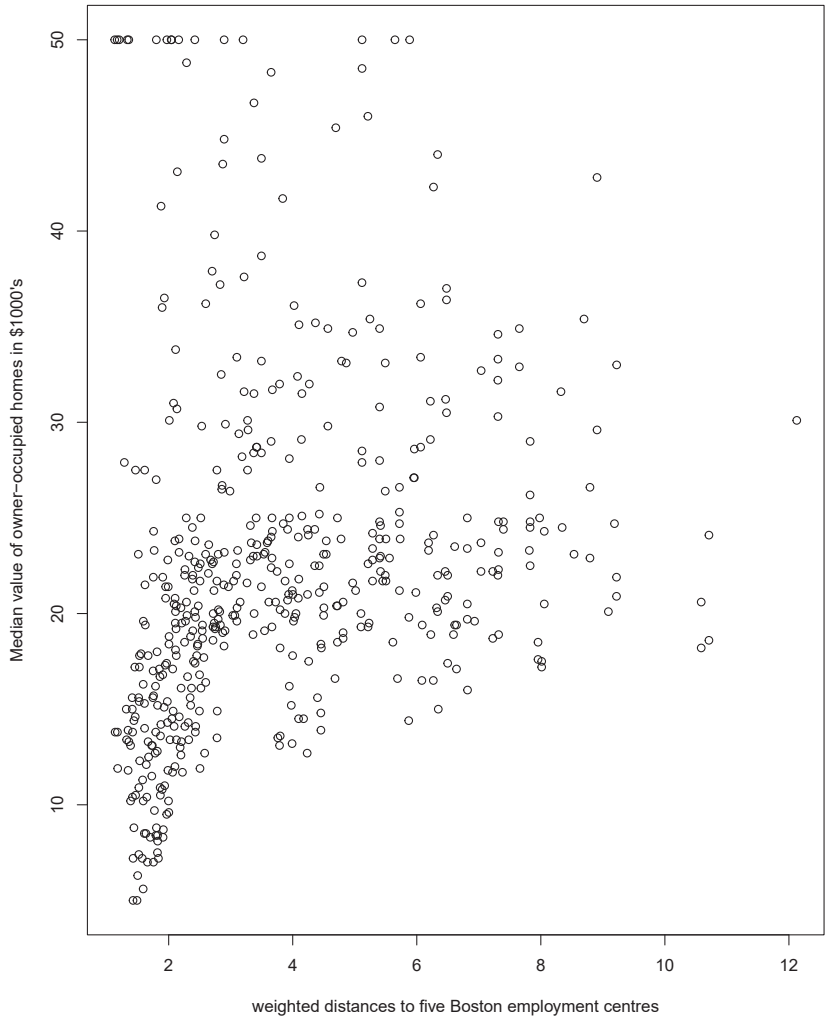


Fig. 1 Scatterplot of Median income and weighted distances to five Boston employment centres

non-additive underlying models, while the local polynomial regression had performance much closer to that of the additive regression for additive models. The additive regression did not improve with increasing sample sizes when the model was misspecified, while the local polynomial regression always improved the results with larger samples.

The variable selection procedures using additive regression and local polynomial estimators, which rely on the fit studied show that the the local polynomial estimator achieves better results. The average number of insignificant selected predictors is always 0 for the local polynomial method, while in average 22% or more of the time the additive model selects unwanted predictors.

## References

[1] R. L. Eubank, *Nonparametric Regression and Spline Smoothing*, Statistics: A Series of Textbooks and Monographs, 1999.

[2] P. J. Green and B.W. Silverman, *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*, Chapman & Hall, 1994.

[3] S. Efromovich, *Nonparametric Curve Estimation: Methods, Theory, and Applications*, Springer Series in Statistics, 1999.

[4] D. Ruppert, M. P. Wand, U. Holst and O. Hssjer, *Local Polynomial Variance-Function Estimation*, Technometrics, 39, pp. 262-273, 1997.

[5] R. T. Rust, *Flexible Regression*, Journal of Marketing Research, 25, pp. 10-24, 1988.

[6] S. Durrleman and R. Simon, *Flexible regression models with cubic splines*, 8, pp. 551-561, 1989.

[7] C. J. Stone, *Additive Regression and Other Nonparametric Models*,The Annals of Statistics, 13, pp. 689-705, 1985.

[8] D. L. Donoho, *High-dimensional data analysis: The curses and blessings of dimensionality*, AMS Conference on Math and Challenges of the 21st Century.

[9] W. Hardle and E. Mammen, *Comparing Nonparametric Versus Parametric Regression Fits*, The Annals of Statistics, 21, 1926-1947, 1993.

[10] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd Edition, Wiley.

[11] G. A. Davis and N. L. Nihan, *Nonparametric Regression and Short-Term Freeway Traffic Forecasting*, Journal of Transportation Engineering, 117, 1991.

[12] J. G. Staniswalis and J.J. Lee, *Nonparametric Regression Analysis of Longitudinal Data*, Journal of the American Statistical Association, 93, pp. 1403-1418, 1998.

[13] P. Constans and J.D. Hirst, *Nonparametric Regression Applied to Quantitative Structure Activity Relationships*, Journal of Chemical Information and Modeling, 40, pp 452-459, 2000.

[14] J. Qiu, H. Wang, D. Lin and B. He, *Nonparametric regression-based failure rate model for electric power equipment using lifecycle data*,

Transmission and Distribution Conference and Exposition (T&D), 2016 IEEE/PES.
[15] E. A. Nadaraya, *On Estimating Regression*, Theory of Probability and its Applications, 9, pp. 141-142, 1964.
[16] G. S. Watson, *Smooth regression analysis*, Sankhya: The Indian Journal of Statistics, Series A, 26, 359-372, 1964.
[17] W. S. Cleveland, *Robust Locally Weighted Regression and Smoothing Scatterplots*, Journal of the American Statistical Association, 74, 829-836, 1979.
[18] W. S. Cleveland, *LOWESS: A program for smoothing scatterplots by robust locally weighted regression*, The American Statistician, 35, 1981.
[19] M. P. Wand and M.C Jones, *Kernel Smoothing*, Chapman & Hall, 1995.
[20] Fan, J. and Gijbels, I, *Local Polynomial Modelling and its Applications*, Boca Raton: Chapman and Hall, 1996.
[21] E. Masry, *Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates*, Journal of Time Series Analysis, 17, pp. 571-599, 1996.
[22] D. Ruppert and M.P.. Wand, *Multivariate Locally Weighted Least Squares Regression*, The Annals of Statistics, 22, pp. 1346-1370, 1994.
[23] J. H. Friedman and W. Stuetzle, *Projection Pursuit Regression*, Journal of the American Statistical Association, 76, 817-823, 1981.
[24] T. J. Hastie and R.J. Tibshirani, *Generalized Additive Models*, Chapman & Hall, 1990.
[25] A. Buja, T. Hastie and R.Tibshirani, *Linear Smoothers and Additive Models*, The Annals of Statistics, 17, 453-555, 1989.
[26] J.D. Opsomer, *Asymptotic Properties of Backfitting Estimators*, Journal of Multivariate Analysis, 73, 166-179, 2000.
[27] C. M. Hurvich, J. S. Simonoff and C.-L. Tsai, *Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion*, Journal of the Royal Statistical Society. Series B, 60, pp. 271-293, 1998.
[28] Boston Housing Dataset, available at https://archive.ics.uci.edu/ml/datasets/Housing.

**Adriano Zanin Zambom** received his Ph.D. in Statistics from Penn State University, studying hypothesis testing and variable selection in nonparametric regression. Soon after receiving his Ph.D., he became a post-doctoral scholar at the State University of Campinas, Brazil, where he later became an Assistant Professor. He taught at Penn State for one semester before starting his career in the Department of Mathematics and Statistics at Loyola University Chicago. His current research involves applied techniques to time series analysis, spatial statistics and trajectory estimation.

**Preethi Ravikumar** received her Master's degree in statistics from the Department of Mathematics and Statistics at Loyola University Chicago under the supervision of Dr. Zambom. She is currently a statistician at a biopharmaceutical company in Seattle, WA.