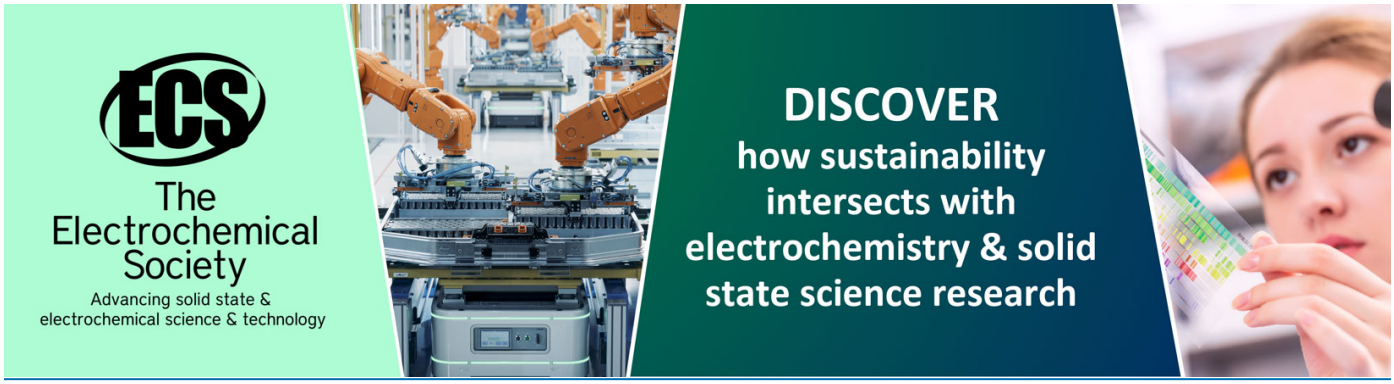
An estimation procedure with knot selection for multivariate regression splines



An estimation procedure with knot selection for multivariate regression splines

Tzee-Ming Huang[1,*]

[1]National Chengchi University, Department of Statistics,16605 Zhinan Road, Taipei, Taiwan (R.O.C.)

*E-mail: tmhuang@nccu.edu.tw

Abstract. Knot selection for regression splines is crucial to the approximation power of splines. For univariate splines, many knot selection algorithms are available. However, extending those univariate algorithms to the multivariate case can be challenging. In this work, an estimation procedure with knot selection for multivariate regression splines is proposed, where the knot selection part is adapted from an existing univariate knot selection algorithm. Simulation results are included to demonstrate the performance of the proposed method.

1. Introduction

In nonparametric regression, one common approach for estimating the regression function is to approximate the regression function using known basis functions and then estimate the coefficients of basis functions using the method of least squares. Some examples of basis functions that have good approximation properties and have been used in nonparametric regression are wavelets [1] and spline functions [2]. Approximation based on fuzzy transforms is also a tool for function approximation (e.g. [3] and [4]). However, to apply this tool in nonparametric regression, an estimation procedure needs to be proposed and its statistical properties are to be investigated. In this paper, the focus is on nonparametric regression using spline approximation with least squares estimation.

In nonparametric regression, when the regression function is approximated using a spline function, the knot locations of the spline function have a significant effect on the approximation accuracy. There are different approaches for knot determination for regression splines. For instance, one can use equally spaced knots and determine the number of knots based on criteria such as the Bayesian information criterion (BIC) or Akaike information criterion (AIC). When using equally spaced knots splines to approximate a smooth regression function to obtain a regression function estimator, Stone [5] showed that the rate of convergence is optimal in the sense of [2] if the number of knots grows at a proper rate as the sample size increases.

In addition to the equally-spaced knot approach, another approach is to estimate the knot locations as unknown parameters, such as in [6] and [7]. While this approach offers great flexibility for determining knot locations, it is often time consuming to search for the best knot locations. When the regression function is multivariate, it is expected that the optimization of knot locations become even more challenging.

An alternative to the knot estimation approach is to perform knot selection. See [8] and [9] for example. Huang [10] proposed an algorithm for knot selection for univariate regression splines based on statistical testing. In this work, an extension of Huang's algorithm to the multivariate case is

considered, and an estimation procedure based on the algorithm is proposed. The rest of the paper is organized as follows. The proposed estimation procedure for multivariate regression splines is given Section 2. Results of simulation experiments are given in Section 3. Conclusions are given in Section 4.

2. Methodology

In this section, the problem set-up will be described and the proposed estimation procedure will be introduced. In addition, another estimation method that is compared with the proposed method in Section 3, will also be introduced. Suppose that $(\boldsymbol{X}_i, Y_i): i = 1,\ldots,n$ are n independent observations from the following nonparametric regression model:

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, i = 1,\ldots,n, \tag{1}$$

where f is the regression function to be estimated and ε_i s are errors. Let d be the dimension of each $\boldsymbol{X}_i$. The knot selection problem arise when one tries to approximate f using a spline function. For the univariate case where $d = 1$, if the regression function f in (1) can be approximated well using a spline of order m and knots $\xi_1,\ldots,\xi_k$. Then an approximate model for (1) is

$$Y_i = \sum_{j=0}^{m-1} \alpha_j p_j(X_i) + \sum_{j=1}^{k} \beta_j B_{m-1,\xi_j}(X_i) + \varepsilon_i, i = 1,...,n, \tag{2}$$

where α_j s and β_j s are coefficients, $p_j(x) = x^j$ for $j = 0,\ldots,m-1$ and for a real number ξ,

$$B_{m-1,\xi}(x) = \begin{cases} (x-\xi)^{m-1} & \text{if } x > \xi; \\ 0 & \text{otherwise.} \end{cases}$$

For the multivariate case, one may use the tensor product of univariate basis functions for splines as multivariate basis functions. However, if all tensor basis functions are used, the number of basis functions can be quite large, especially when many knots are used for each dimension. The tensor basis functions used in the proposed method only include basis functions of one the two forms: $B_{a_1,\ldots,a_d}$, and $p_{j_1,\ldots,j_d}$, where for real numbers $a_1,\ldots,a_d$ and integers $j_1,\ldots,j_d \in \{0,\ldots,3\}$,

$$B_{a_1,\ldots a_d}(x_1,\ldots,x_d) = \prod_{l=1}^{d} B_{3,a_l}(x_l)$$

and

$$p_{j_1,\ldots j_d}(x_1,\ldots,x_d) = \prod_{l=1}^{d} p_{j_l}(x_l)$$

for real numbers $x_1,\ldots,x_d$. For the proposed method, the basis functions are selected from a collection of basis functions determined by some knot vectors $\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_k$, where $\boldsymbol{\xi}_j = (\xi_{j,1},\ldots,\xi_{j,d})$ is a d-dimensional vector for $j = 1,\ldots,k$. Let $S_{\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_k}$ denote the collection of basis functions determined by knot vectors $\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_k$, then

$$S_{\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_k} = S_0 \cup \left\{ B_{\xi_{j,1},\ldots,\xi_{j,d}} : j \in \{1,\ldots,k\} \right\} \tag{3}$$

where $S_0 = \left\{ p_{j_1,\ldots,j_d} : j_1,\ldots,j_d \in \{0,\ldots,3\} \right\}$. Thus the approximation model for (1) considered in this study is

$$Y_i = \sum_j \alpha_j B_j^*(\boldsymbol{X}_i) + \varepsilon_i, i = 1,\ldots,n, \tag{4}$$

where B_j^* s are some basis functions selected from $S_{\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_k}$ and α_j s are the coefficients. Next, we will give the details of the proposed estimation procedure in Section 2.1.

2.1. The proposed estimation procedure and the BIC-Backward algorithm

As mentioned before, for the proposed method, the basis functions are selected from $S_{\xi_1,\ldots,\xi_k}$, which is defined in (3), where $\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_k$ are knot vectors of dimension . In this section, we will give details of how to select the knot vectors.

To determine whether a knot vector $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_d)$ is needed, we construct a random test $\psi_{0.05,\delta,\boldsymbol{\xi}}$, which is an extension of the test in [10]. The test $\psi_{0.05,\delta,\boldsymbol{\xi}}$ is based on $(\boldsymbol{X}_i, Y_i)$ s such that $\| \boldsymbol{X}_i - \boldsymbol{\xi} \| < \delta$, where $\| \cdot \|$ is the Euclidean norm. To obtain the test statistic, we generate IID coefficients $(b_1,\ldots,b_d)$ from the uniform distribution on [−1,1]. Let $\boldsymbol{b} = (b_1,\ldots,b_d)^T$ be the vector of the generated coefficients. The hyperplane $\{\boldsymbol{x} : \boldsymbol{b}^T(\boldsymbol{x}-\boldsymbol{\xi}) = 0\}$ splits the $\boldsymbol{X}_i$ s such that $\| \boldsymbol{X}_i - \boldsymbol{\xi} \| < \delta$ into two groups $\{\boldsymbol{X}_i : i \in \Lambda_1\}$ and $\{\boldsymbol{X}_i : i \in \Lambda_2\}$, where

$$\Lambda_1 = \{i \in \{1,\ldots,n\} : \| \boldsymbol{X}_i - \boldsymbol{\xi} \| < \delta \text{ and } \boldsymbol{b}^T(\boldsymbol{X}_i - \boldsymbol{\xi}) > 0\}$$

and

$$\Lambda_2 = \{i \in \{1,\ldots,n\} : \| \boldsymbol{X}_i - \boldsymbol{\xi} \| < \delta \text{ and } \boldsymbol{b}^T(\boldsymbol{X}_i - \boldsymbol{\xi}) < 0\}.$$

For j = 1, 2, let n_j be the number of elements in Λ_j. If both n_1 and n_2 are at least 40, regress $\{Y_i\}_{i\in\Lambda_j}$ on tensor products of univariate linear functions of $\{\boldsymbol{X}_i\}_{i\in\Lambda_j}$ using linear regression and let $\hat{a}_j$ be the $2^d \times 1$ least squared estimator of regression coefficient vector and RSS_j be the residual sum of squares. Then, compute the conditional covariance matrix of $\hat{a}_1 - \hat{a}_2$ given $\boldsymbol{X}_i$ s: $i \in \Lambda_1 \cup \Lambda_2$ assuming that the regression errors are ε_i s, and ε_i s are IID $N(0, \sigma^2)$ and are independent of $\boldsymbol{X}_i$ s. Under the above assumption, the conditional covariance matrix of $\hat{a}_1 - \hat{a}_2$ given $\boldsymbol{X}_i$ s: $i \in \Lambda_1 \cup \Lambda_2$ is σ^2 multiplied by a known matrix. Let Σ denote the matrix such that $\sigma^2\Sigma$ is the conditional covariance matrix of $\hat{a}_1 - \hat{a}_2$ given $\boldsymbol{X}_i$ s: $i \in \Lambda_1 \cup \Lambda_2$. Then the test statistic is given by

$$W_{\boldsymbol{b}} \overset{def}{=} \frac{(\hat{a}_1 - \hat{a}_2)^T \Sigma^{-1} (\hat{a}_1 - \hat{a}_2)}{\hat{\sigma}^2}, \tag{5}$$

where

$$\hat{\sigma}^2 = \frac{RSS_1 + RSS_2}{n_1 + n_2 - 2}.$$

The above process of generating $\boldsymbol{b}$ to obtain $W_{\boldsymbol{b}}$ is repeated until three $W_{\boldsymbol{b}}$ s are obtained. However, the number of repetitions is limited to 10. Let W be the maximum of the $W_{\boldsymbol{b}}$ s. The maximum is defined as $-\infty$ if there is not any $W_{\boldsymbol{b}}$ obtained. Then W is the test statistic for $\psi_{0.05,\delta,\boldsymbol{\xi}}$. If W is greater than the 95% quantile of the χ^2 distribution with 2^d degrees of freedom, then $\boldsymbol{\xi}$ is considered as an important knot vector. The proposed estimation procedure for f is given below.

Algorithm 2.

- For $j = 1, \ldots, 5$, take $\delta = 1/2^j$ and carry out Steps (i) and (ii).
  - (i) For $i = 1, \ldots, n$, conduct the test $\psi_{0.05,\delta,\boldsymbol{X}_i}$. Select $\boldsymbol{X}_i$ as a knot vector if the p-value is less than 0.05.
  - (ii) Let T_j be the set of selected knot vectors from (i). Take the regression model in (4) as the full model and then perform backward elimination of variables to obtain a sequence of reduced models. Choose the reduced model with the smallest BIC value. Let BIC_j denote the smallest BIC value.

- Let BIC_{j_0} be the smallest BIC value among $BIC_1, \ldots, BIC_5$. The estimated regression function in the model corresponding to BIC_{j_0} is the final estimator of the regression function f.

For comparison purpose, another estimation method based on equally-spaced knots with backward elimination is considered. The method will be referred as BIC-Backward. The details of BIC-Backward are given below.

Algorithm for BIC-Backward.

- For $j = 1, \ldots, 4$, take $\delta = 1/2^j$. Let $(\xi_1, \ldots, \xi_k) = (\delta, 2\delta \ldots, 1-\delta)$ be the k equally spaced knot points in (0, 1) with $k = 2^j - 1$. Let $S_{full,\delta}$ be the set of tensor basis functions, where for each dimension, the univariate basis functions are $p_0, \ldots, p_3, B_{3,\xi_1}, \ldots, B_{3,\xi_k}$.
  - (i) Take the regression model in (4) as the full model with the basis functions B_j^*s replaced by the basis functions in $S_{full,\delta}$ and then perform backward elimination of variables to obtain a sequence of reduced models. Choose the reduced model with the smallest BIC value. Let BIC_j denote the smallest BIC value.
- Let BIC_{j_0} be the smallest BIC value among $BIC_1, \ldots, BIC_5$. The estimated regression function in the model corresponding to BIC_{j_0} is the final estimator of the regression function f.

Note that $j = 5$ is not used in the algorithm for BIC-Backward since the number of basis functions exceeds the sample size $n = 1000$ in such case.

3. Simulation results

A simulation experiment has been carried out to examine the performance of the proposed estimation procedure and the results are given in this section.

For the simulation experiment, the data generating process is described below. 100 data sets are generated from model (1) with $n = 1000$, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are IID random vectors from the uniform distribution on $(0,1)^2$, ε_i s are IID random variables from $N(0, \sigma^2)$ with various σ values, and $f \in \{f_1, f_2, f_3\}$. Here f_1 is the probability density function for the bivariate normal distribution with mean (0.4, 0.4) and diagonal covariance matrix with diagonal vector (0.05, 0.05),

$$f_2(u,v) = \begin{cases} c(u/2, v/2) & \text{if } (u,v) \in [0.2, 0.8]^2; \\ 0 & \text{otherwise.} \end{cases}$$

where c is the density function of the Joe copula with parameter 40, and f_3 is the two dimensional AMH copula density with parameter −0.8. The cumulative distribution function of the Joe copula with parameter θ is the function C_1 given by

$$C_1(u,v) = 1\text{-}((1\text{-}u)^\theta + (1-v)^\theta - (1-u)^\theta (1-v)^\theta)^{1/\theta}$$

for $(u,v) \in (0,1)^2$. The cumulative distribution function of the AMH copula with parameter is the function C_2 defined by

$$C_2(u,v) = \frac{uv}{1-\theta(1-u)(1-v)}$$

for $(u,v) \in (0,1)^2$. The graphs of f_1, f_2, f_3 on $(0,1)^2$ are shown in the left, middle and right pannels of figure 1 respectively.

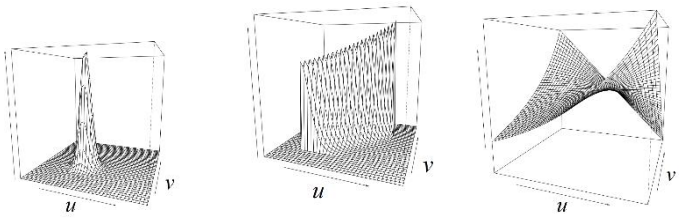


Figure 1. Graphs of f_1, f_2, f_3.

For each of the generated data set, the proposed method (Algorithm 2) is applied to obtain an estimated regression function $\hat{f}$. The mean squared error

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))^2$$

is computed to evaluate the estimation accuracy. For comparison purpose, the algorithm for BIC-Backward is also applied to generated data sets.

For each of the three functions f_1, f_2, f_3 and for each σ, the mean and standard deviation of the 100 mean squared errors are reported in table 1 for the BIC-Backward method and the proposed method. The σ values are chosen so that the ratio

$$r=\frac{E\,|f(\boldsymbol{X})|}{\sigma}$$

is about 140 or 7 when the distribution of $\boldsymbol{X}$ is the uniform distribution on $(0,1)^2$.

From the results in table 1, for f_1 and f_2, which need to be approximated by splines with many knots, the proposed method performs better than the BIC-Backward method when σ is small, and the BIC-Backward method performs better than the proposed method when σ is large. For f_3, which is a smooth function, the proposed method performs better for the large σ case, and the BIC-Backward method performs better for the small σ case.

Table 1. Averages and standard deviations for mean squared errors.

	σ	BIC-Backward	Algorithm 2
f_1	3.890×10^{-2}	7.695×10^{-5} (2.629×10^{-5})	2.989×10^{-5} (1.023×10^{-5})
f_1	7.781×10^{-1}	2.316×10^{-1} (3.644×10^{-2})	2.911×10^{-1} (2.880×10^{-2})
f_2	2.117×10^{-2}	1.697×10^{-4} (8.929×10^{-5})	1.463×10^{-4} (7.026×10^{-5})
f_2	4.234×10^{-1}	2.2189×10^{-1} (8.949×10^{-2})	2.527×10^{-1} (6.812×10^{-2})
f_3	1.636×10^{-3}	1.322×10^{-7} (2.608×10^{-8})	1.271×10^{-6} (1.562×10^{-7})
f_3	3.272×10^{-2}	2.2534×10^{-5} (7.555×10^{-6})	2.100×10^{-5} (6.142×10^{-6})

4. Discussion and conclusions

In this article, an estimation procedure for nonparametric regression is proposed. The proposed procedure is based on spline approximation and a knot vector selection step is involved. The knot vector selection step is adapted from the univariate knot selection algorithm in [10]. One major modification is that the requirement that the distance between two knots is at least δ is removed so that knot vectors that are close to one another may be selected when necessary. To avoid the overfitting

problem when too many knot vectors are selected, the backward elimination step is included in the proposed procedure.

References

[1] David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. Ann. Statist., 26(3):879–921, 1998.
[2] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. Ann. Statist., 10(4):1040–53, 1982.
[3] Irina Perfilieva. Fuzzy transforms: Theory and applications. Fuzzy Sets and Systems, 157(8):993–1023, 2006.
[4] Martins Kokainis and Svetlana Asmuss. Approximation by multivariate higher degree F-transform based on B-splines. Soft Computing, 21(13):3587–614, 2017.
[5] Charles J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. Ann. Statist., 22(1):118–71, 1994.
[6] Mary J. Lindstrom. Penalized estimation of free-knot splines. Journal of Computational and Graphical Statistics, 8(2):333–52, 1999.
[7] Shanggang Zhou and Xiaotong Shen. Spatially adaptive regression splines and accurate knot selection schemes. Journal of the American Statistical Association, 96:247–59, 2001.
[8] Yuan Yuan, Nan Chen, and Shiyu Zhou. Adaptive B-spline knot selection using multi-resolution basis set. IIE Transactions, 45(12):1263–77, 2013.
[9] V. K. Kaishev, D. S. Dimitrova, S. Haberman, and R. J. Verrall. Geometrically designed, variable knot regression splines. Computational Statistics, 31(3):1079–105, 2016.
[10] Tzee-Ming Huang. A knot selection algorithm for regression splines. In Proceedings of the 62th ISI World Statistics Congress (August 2019), To appear.