응용통계연구 제 5권 2호
1992년 pp 181-192

A Study on Selection of Tensor Spline Models[1]

Ja-Yong Koo[2]

ABSTRACT

We consider the estimation of the regression surface in generalized linear models based on tensor-product B-splines in a data-dependent way. Our approach is to use maximum likelihood method to estimate the regression function by a function from a space of tensor-product B-splines that have a finite number of knots and are linear in the tails. The knots are placed at selected order statistics of each coordinate of the sample data. The number of knots is determined by minimizing a variant of AIC. A numerical example is used to illustrate the performance of the tensor spline estimates.

1. Introduction

Let Y be a response variable whose distribution depends on the values of a vector of covariates $x=(x_1,\cdots,x_d)$ and let $\eta=f(x)$ denote a parameter defined in terms of this distribution. If η is the mean of Y, f is the usual regression function. Suppose instead that Y takes on only the values 0 and 1 and that η is the logit of the probability that Y=1, where $\text{logit}(\pi) = \log(\pi/(1-\pi))$. Then f is the logistic function.

Nelder and Wedderburn (1972) introduced generalized linear models (GLMs) including the above two examples. In GLMs we assume a response variable Y for which the likelihood can be written in the form

$$l(\theta;y)=\{y\theta-b(\theta)\}/a(\phi)+c(y,\ \phi) \tag{1.1}$$

where θ is the canonical parameter and $a(\phi)$ is the scale parameter. The mean μ is related to covariates $x=(x_1,\ldots,x_d)$ via the link function g such

[1] This research was supported by Korea Research Foundation

[2] Department of Statistics, Hallym University, Chunchon 200-702, Korea.

that $\eta = g(\mu) = f(x)$. The GLMs assume that the regression surface η is linear in covariates, i.e $\eta = \sum \beta_j x_j$ and the parameters β_j are estimated using the maximum likelihood method if appropriate distributional assumption is made; otherwise the procedure is justified on the basis of quasi-likelihood. However if the functional form of the dependence of predictor η on the covariates is uncertain, it is important to have some methods available to estimate the regression surface f(x) nonparametrically.

Koo and Lee(1992) studied tensor-product B-splines for GLMs and compared some previously proposed nonparametric alternatives, such as generalized additive models of Hastie and Tibshirani(1986), penalized likelihood alternative of O'Sullivan, et. al.(1986) and MARS of Friedman (1991).

In this paper, we attempt to combine tensor-product B-spline method for GLMs by Koo and Lee(1992) and Smith's(1982) automated knot selection procedure to estimate the regression surfaces $\eta = f(x)$ in GLMs. The tensor-product spline estimate of regression surface f(x) can be obtained by directly maximizing the log likelihood over the space of the tensor-product splines. This method is particularly suited to the analysis of larger data sets (say n>50 data points) as is multivariate smoothing splines of O'Sullivan, et. al. (1986). From the simulation study we have found that knots remains where there is high local curvature and knots are deleted if there is no such features, which means that tensor spline estimates with knot deletion algorithm has the local adaptivity.

2. Tensor spline estimates in GLMs

The spline is an attractive tool for nonparametric function estimation. A spline of degree q is a piecewise q-th degree polynomial, perhaps subject to some smoothness constraints at the knots(boundaries between consecutive pieces). Commonly employed are piecewise constants (q=0), linear splines (q=1), quadratic splines (q=2) and, especially, cubic splines (q=3). In practice twice continuously differentiable cubic splines are particularly attractive since modest discontinuities in the third derivative cannot be detected visually. It has been suggested that the spline functions should be restricted linear at each tail. Fuller(1969) proposed such linear restrictions in the context of extrapolating a time series trend. See also Stone and Koo(1986). If we adopt cubic splines restricted to have each tail linear, the splines satisfy the natural boundary condition that their higher derivatives other than first derivative vanish. Even when the domains of covariates are unbounded, we recommend the use of such splines.

For the simplicity of presentation, we consider GLMs with bivariate covariate. Given N_j knots for each covariate x_j and $j=1,2$, let S_j be the space of cubic splines of argument x_j with linear constraint. Then S_j is a N_j-dimensional vector space whose basis $\{B_{j,k}:1\leq k\leq N_j\}$ can be constructed starting from either a truncated power basis or a B-spline basis [see de Boor(1978)]. The element of the bivariate tensor-product B-spline space T with linear constraint is a tensor product of S_1 and S_2 whose elements can be represented as

$$\sum_i \sum_j \beta_{ij} B_{1,j}(x_1) B_{2,j}(x_2) \ .$$

The tensor-product spline basis for T, $A_k(x) = B_{1,i}(x_1) \cdot B_{2,j}(x_2)$, is represented by the product of basis of each component and hence the dimension of T is given by $J=N_1 \cdot N_2$. It has a major computational advantage in that the properties of B-splines in one dimension carry over to the bivariate tensor-product B-splines. See Schumaker(1981) for the property of tensor-product B-splines.

Now let (X_i,Y_i) for $i=1,\cdots,n$ be independently and identically distributed random variables, where $X_i=(X_{i1},X_{i2})$. Suppose that the conditional distribution of Y_i given $X_i=x_i$ belongs to the GLM family of distributions but the form of the regression surface $\eta_i=f(x_{i1},x_{i2})$ is unknown. Since we do not restrict the functional form of f, the choice of link function is not critical and thus we assume the canonical link $\theta=\eta$. Given the number and location of knots, we approximate the regression surface f by a tensor-product spline $s(x;\beta)=\sum \beta_k A_k(x)$ in T. Then the log-likelihood is proportional to

$$l(\beta)=\sum\{s(x_i;\beta)y_i-b(s(x_i;\beta))\}/a(\phi), \qquad (2.1)$$

where $x_i=(x_{i1},x_{i2})$. The tensor spline regression surface estimator $\hat{f}(x)=\sum\hat{\beta}_k A_k(x)$ can be obtained by directly maximizing the log-likelihood (2.1). To this end we use the Newton-Raphson method which is equivalent to an iteratively reweighted regression procedure. For the log-likelihood (2.1) the Fisher scoring method is equivalent to the Newton-Raphson. Since the likelihood function $l(\beta)$ is concave in β provided var(Y) >0, the existence and uniqueness of MLE of β is obvious. For details see Koo and Lee(1992). Programs developed by de Boor(1978) are used with slight modification to implement this procedure.

3. Stepwise knot deletion

In order to implement tensor-product splines, the rules for selecting the number and location of knots for each covariate must be considered. Choosing the number of knots is very important problem: it is comparable to choosing a bandwidth in kernel regression estimation or smoothing parameter in penalized likelihood method of O'Sullivan, et. al. (1986). Too many knots leads to a noisy estimate; too few knots gives an estimate that is overly smoothed and thereby missing essential details. The problem of knot placement is also important, since many knots are needed where there is high local curvature of f and few knots may be enough where the fluctuation of f is mild. An algorithm achieving these two goals is this [see also Smith(1982)] : start out with a larger number of knots and then remove those knots that appear to be unessential for the given data.

The following lemma details the linear combination of the bivariate tensor-product B-spline which are used to test the importance of breakpoints.

Lemma. Let $s(\cdot;\beta)=\sum_{j=1}^{N_1}\sum_{k=1}^{N_2}\beta_{jk}B_{1,j}(x_1)B_{2,k}(x_2)$ be a bivariate tensor-product spline in T. The absence of an interior breakpoint ξ in the function $s(\cdot;\beta)$ in x_1 direction occurs if and only if the contrasts

$$\sum_j \beta_{jk}\{B_{1,j}^{(p)}(\xi^-)-B_{1,j}^{(p)}(\xi^+)\}$$

for $p=0,1,2,3$ and $k=1,\cdots,N_2$ all equal zero. Here $B_{1,j}^{(p)}(\xi^-)$ and $B_{1,j}^{(p)}(\xi^+)$ are respectively the left- and right-hand limit of $\partial^p B_{1,j}/\partial x_1^p$ at ξ.

Proof. When we consider $s(\cdot;\beta)$ as a function of x_1 alone, it is a polynomial. Thus the absence of an interior breakpoint ξ in x_1 direction means that the polynomial pieces on either side of ξ are the same, i.e., they agree in function value and all derivatives. Since the derivatives with respect to x_1 beyond the second are identically zero, it is necessary and sufficient to require equal function values and p-th derivative values for $p=1,2,3$ from which

$$\sum_k\Big[\sum_j \beta_{jk}\{B_{1,j}^{(p)}(\xi^-)-B_{1,j}^{(p)}(\xi^+)\}\Big]B_{2,k}(x_2)=0 \text{ for all } x_2. \tag{3.1}$$

Since $\{B_{2,k}(x_2),\ k=1,\cdots,N_2\}$ are linearly independent, (3.1) implies the coefficients $\sum_j \beta_{jk}\{B_{1,j}^{(p)}(\xi^-)-B_{1,j}^{(p)}(\xi^+)\}$ for $k=1,\cdots,N_2$ all equal zero. This completes the proof of Lemma.

When ξ is a simple knot and g is twice continuously differeniable, the absence of ξ in x_1 direction occurs if the contrasts $\sum_j \beta_{jk}\,\{\,(B_{1,j}^{(3)}(\xi^-)-B_{1,j}^{(3)}(\xi^+)\,\}$ equal to zero for $k=1,\cdots,N_2$.

The estimates $(\hat{\beta_{jk}})$ can be obtained by the maximum likelihood method as described in Section 2, and let $\Gamma(\hat{\beta})$ be the inverse of the infomation matrix. Let $R(\xi)$ be a $N_2\times J$ matrix such that the k-th element of $R(\xi)\beta$ is

$$\sum_j \beta_{jk}\{B_{1,j}^{(3)}(\xi^-)-B_{1,j}^{(3)}(\xi^+)\},$$

where $J=N_1 \cdot N_2$. Note that the condition in the above Lemma is equivalent to $R(\xi)\beta=0$. By the above Lemma, we can use a test statistics

$$(R(\xi)\hat{\beta})^t\{R(\xi)\Gamma(\hat{\beta})R(\xi)^t\}^{-1}R(\xi)\hat{\beta}/N_2$$

for the hypothesis $R(\xi)\beta=0$, i.e., the absence of an interior breakpoint ξ in the function $s(\cdot;\beta)$ in x_1 direction. This test statistics is proposed from the heuristic point of view by assuming that $\hat{\beta}$ is approximately normal with mean β and covariance matrix $\Gamma(\hat{\beta})$. But the distributional property of this statistic has not been shown.

Consider $t_{j,k}$, $1 \leq k \leq N_j$, $j=1,2$, as being non-permanent initial knots that may be deleted and consider stepwise knot deletion among the non-permanent initial knots. We start the knot deletion procedure with as many knots as the computing time permits. A simple example of initial knot placement is the rule that puts down N_j knots along the x_j-axis as closely as possible to equispaced order statistics of j-th coordinate. See Kooperberg and Stone(1990) for a more complicated rule for knot placement. Also we might be able to estimate the location of knots, which method is not used here because of its computing time. At any step we delete that knot having the smallest value of

$$(R(t_{j,k})\hat{\beta})^t\{R(t_{j,k})\Gamma(\hat{\beta})R(t_{j,k})^t\}^{-1}R(t_{j,k})\beta.$$

In this manner, we arrive at a sequence of models indexed by (m_1,m_2); the (m_1,m_2)-th model has $(N_1-m_1)\cdot(N_2-m_2)$ free parameters. Let $\hat{l}_m$ denote the log-likelihood function for the $m=(m_1,m_2)$-th model evaluated at the maximum-likelihood estimate for that model. Let

$$AIC_{\alpha,m}=-2\,\hat{l}_m+\alpha(N_1-M_1)\cdot(N_2-m_2)$$

be the Akaike Information Criteria with parameter penalty α for the m-th model. Since the traditional value of $\alpha=2$ often leads to spurious models, Kooperberg and Stone(1990) suggested $\alpha=3$ and Schwarz(1978) recommended $\alpha=\log n$. We choose the model corresponding to that value

$\hat{m}$ of m that minimizes $AIC_{3,m}$. Here is a flowchart for stepwise knot deletion algorithm.

(1) Place as many knots as the computing time permits.
(2) Delete knots with the minimum test statistics.
(3) Choose the tensor spline model with minimum $AIC_{3,m}$.

Remark. If we use a criterion with large α, it is better to have a small number of knots so that the estimated regression surface becomes smoother. The smoothing operation prevents unnecessary fluctuations of the surface estimator but may, of course, lose fine structure, especially where there is high local curvature; as ever, due care is necessary in determining the penalty parameter α. A reason of using AIC instead of say, Cross-Validation which is asymptotically equivalent to AIC is that AIC is easier to use, since the maximized log-likelihood is obtained by product. Furthermore, at each deletion step we have to find the estimate $\hat{\beta}$ by an iterative algorithm which prevent us from using a time consuming criteria such as cross-validation.

Example. For each i=1,...,n, suppose that given $X_i=x_i$, the random variable Y_i has the Bernoulli distribution with parameter

$$\pi(x) = Pr(Y_i = 1 \mid X = x) = f_1(x)/(f_1(x) + f_2(x)), \tag{3.2}$$

where

$$f_1(x)=\exp\{-(x_1^2+x_2^2)/2\}$$

and

$$f_2(x)=\exp\{-(x_2+1.5)^2/2\}[\exp\{-(x_1-2.5)^2/2\}+\exp\{-(x_1+2.5)^2/2\}]/2.$$

Then the logistic regression surface is given by $f(x) = \log[\pi(x)/\{1-\pi(x)\}]$. The tensor spline estimates $\hat{f}$ can be obtained by the Newton-Raphson method as in Section 2. The estimate of $\pi(x)$ can also be obtained as $\exp(\hat{f})/(1+\exp(\hat{f}))$. Villalobos and Wahba (1987) have used the same function for their simulation study. For this example, the values of random variables X_1 and X_2 are generated from uniform distribution with range [-4,4] using the IMSL subroutine GGUBS. Figures (a) and (b) give the perspective plot and contour plot of the true probability $\pi(x)$. Figure (c) and (d) give the perspective plot and contour plot of the tensor spline estimate based on the random sample of size n=200 with knot deletion algorithm. These plots are done by commands 'persp' and 'contour' in S-plus; see Becker et al. (1988). We use the simple rule of initial knot placement putting down $N_j=5$ knots along the x_j-axis to equispaced order statistics of j-th coordinate. Presumable we can use a criteria such as AIC for choosing the number of initial knots, which is currently under study theoretically. However, in practice the problem of choosing initial knots doesn't seem crucial since we have only to use sufficiently many knots. Remaining knots are denoted by tick marks on contour plot of estimated surfaces.

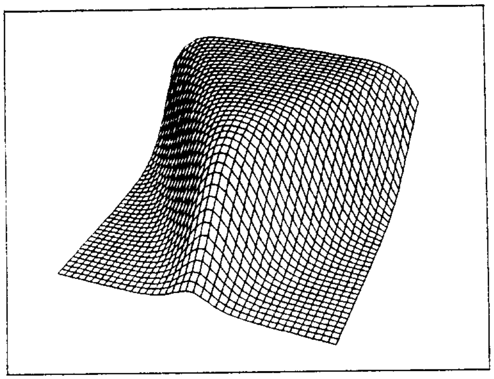

Figure (A)

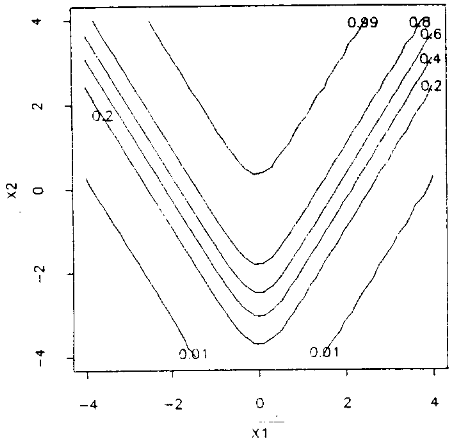

Figure (B)

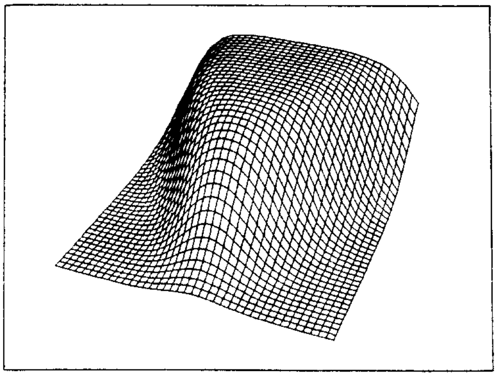

Figure (C)

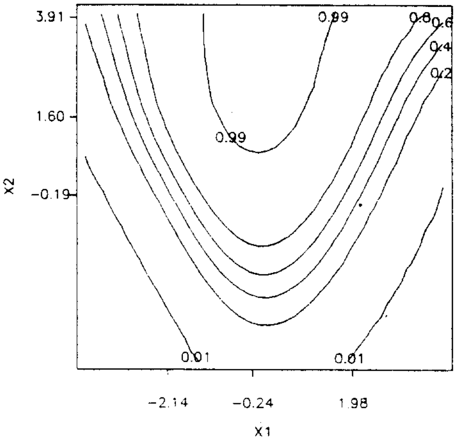

Figure (D)

4. Discussion

In this paper, we illustrate that the stepwise knot deletion is easily implemented for modeling the shape of regression surfaces for complicated data with bivariate covariate, such as those with binomial responses. When the number d of covariates is greater than 2, the dimension J of the space of tensor product splines is given by the product of the dimension of splines of each component, i.e. $J=N_1 \cdots N_d$, which grows rapidly as d increases. If the number of covariates is large, the adaptive tensor spline method given in this paper requires a large amount of data and computing time. By restricting the functional form of the regression surface to the hierarchical model of the form with k small

$$f(x) = f_0 + \sum f_i(x_i) + \sum f_{ij}(x_i, x_j) + \cdots + \sum f_{i_1 \cdots i_k}(x_{i1}, \cdots, x_{ik}),$$

we can avoid such dimensionality problem. However, an obvious drawback of such hierarchical model with k<d is that they cannot estimate f itself, i.e. there remains a model bias. Given a modern workstation environment with fast computation, interactive graphics and hardcopy capability, the low-dimensional hierarchical tensor splines, especially

models having interaction with k=2 which are estimated by bivariate splines, are a promising tool in exploratory statistics.

< References >

[1] de Boor, C.(1978), *A Practical Guide to Splines*. Springer-Verlag, New York.

[2] Becker, R. A., Chambers, J. M. and Wilks, A. R.(1988), *The New S Language*. The Wadsworth & Brooks, CA.

[3] Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines (with discussion)," *Annals of Statististics*, Vol.19, 1-141.

[4] Fuller, W. A. (1969),"Grafted Polynomials as Approximating Functions," *Austrailian Journal of Agricultural Economics.*, Vol. 13, 35-46.

[5] Hastie, T. J. and Tibshirani, R. J. (1986), "Generalized Additive Models," *Statistical Science*, Vol. 1, 297-318.

[6] Koo, J. A. and Lee, Y. J. (1992), "Tensor-product B-Splines in Generalized Linear Models," Manuscript.

[7] Kooperberg, C. and Stone, C. J. (1990), "A Study on Logspline Density Estimation," Technical report #238, Univ. of Calf, Berkeley.

[8] Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society.*, Series. A., Vol. 135, 370-84.

[9] O'sullivan, F., Yandell, B. S. and Raynor, W. J. (1986), "Automatic Smoothing of Regression Functions in Generalized Linear Models," *Journal of the American Statistical Association,* Vol. 81, 96-103.

[10] Schumaker, L. L (1981), *Spline Functions : Basic Theory*. Wiley, New York.

[11] Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annanls of Statistics,* Vol. 6, 461-464.

[12] Smith, P. (1982), "Curve Fitting and Modeling with Splines using Statistical Variable Selection Methods," NASA, Langley Research Center, Employ, VA, NASA Report 166034.

[13] Stone, C. J. and Koo, C. Y. (1986), "Additive Splines in Statistics," In *1985 Statistical Computing Section. Proceedings of the American Statistical Association* 45-48 American Statistical Association,

Washington.

[14] Villalobos, M. and Wahba, G. (1987), "Inequality Constrained Multivariate Smoothing Splines with Application to the Estimation of posterior probabilities," ***Journal of the American Statistical Association.*** Vol. 82, 239-248.

텐서 스플라인 모형 선택에 관한 연구[1]

< 요 약 >

구 자 용[2]

본 논문에서는 텐서 스플라인을 이용하여, 일반화된 선형모형의 회귀함수를 자료에만 의존하는 방식으로 추정하는 문제를 고려하였다. 최우 추정법을 이용하여 회귀 함수를 추정하는데, 이용된 텐서 스플라인은 접목점의 수가 유한개이며, 독립변수 영역의 주변에서는 선형으로 제한되었다. 접목점은 자료의 각 좌표의 순서 통계량에 위치하도록 했고 그 수는 AIC의 변형된 식을 최소로 하는 수로 결정 했다. 모의 실험 예를 통하여 추정량을 예시하였다.

[1] 이 연구는 1990년도 문교부지원 한국 학술진흥재단 지방대육성 학술연구조성비에 의해 연구되었음.

[2] 200-702 강원도 춘천시 옥천동 한림대학교 통계학과