Knot-Placement to Avoid Over Fitting in B-Spline Scedastic Smoothing

Hirokazu Yanagihara* and Megu Ohtaki**

* *Department of Mathematics, Faculty of Science,*
Hiroshima University, Higashi-Hiroshima, 739-8529, Japan

** *Department of Environmetrics and Biometrics,*
Research Institute for Radiation Biology and Medicine,
Hiroshima University, Hiroshima, 734-8553, Japan

AMS 1991 *subject classifications.* Primary 62-04 ; Secondary 62G08.

Key words and Phrases: B-spline smoothing ; Equidistant arrangement ; Equipotent arrangement ; Information criterion ; Knot-placement ; Non-parametric regression ; Optimization algorithm ; Over fitting ; Smoothing parameter.

Abstract

This paper deals with knot-placement in B-spline scedasitic smoothing. We commonly fit B-spline using knots between which each interval is the same width. The number of knots is selected by applying an information criterion with an optimization algorithm. In order to avoid over fitting with this method, we consider other arrangements based on the number of sample points in each interval. We choose knots such that each interval contains about the same number of samples. These knot placement schemes are compared through the Monte Carlo simulation.

1. Introduction

B-spline smoothing of data with errors and a complicated trend is useful, because its computation is simpler than that of other nonparametric methods. In B-spline smoothing, knot selection, which defines the basis functions, is an important problem, because different knot arrangements can result in different smoothing curves. The simplest arrangement is that each interval between knots has the same width. However, some such intervals will include few data points, resulting in a poor estimator of the curve in those intervals. As a result, there is the possibility of over fitting. Other methods of knot-selection may be considered, such as regarding the knots as variables and deriving them by nonlinear least squares optimization (De Boor and Rice (1968)). However, it takes substantial computing time to obtain such knots, and the optimal values might not be obtained because this equation has many polar values. The successive division method (Ichida, Yoshimoto and Kiyono (1976)) requires less computing time, but is difficult to apply because it does not easily accommodate a smoothing parameter. The main purpose of this paper is to study strategies for placing knots to avoid over fitting in automated knot-selection algorithms. In considering this problem, we examine which knot-placement strategies can lead to over fitting and which are computationally most simple and efficient. The strategies we consider are as follows.

(*i*) *equidistant*: each interval between knots is the same width ;

(*ii*) *equipotent*: each interval between knots includes, to the extend possible, the same number of data points ;

(*iii*) *modified equipotent*: a width is guaranteed by removing knots from an equipotent arrangement.

Atilgan and Bozdogan (1990) and Imoto and Konishi (1999a), (1999b) described an algorithm for choosing the number of knots. We use the SPIDER algorithm (Ohtaki and Izumi (1999)) to obtain the smoothing parameter by minimizing an information criterion.

The present paper is organized in the following way. In Section 2, we introduce the B-spline nonlinear regression model. In Section 3, we consider the three knot-placement strategies and study their influences through simulation. The strategies are compared by iterating numerical examples in Section 4.

2. B-spline nonlinear regression model

Let $\{(x_i, y_i) \mid i = 1, 2, \ldots, n\}$ be n observable data pairs on an explanatory variable X and a response variable Y. Consider a regression model

$$y_i = \mu_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

In this model, it is assumed that the ε_i are independently distributed according to a normal distribution with mean 0 and variance σ^2. A nonparametric regression model applies a complicated nonlinear structure to data. Using a smooth function $g(\cdot)$, the model is defined by

$$\mathrm{E}(Y_i \mid x_i) = \mu_i = g(x_i), \quad i = 1, 2, \ldots, n.$$

In the B-spline nonlinear regression model, μ is regarded as a linear combination of known basis functions. That is to say, making use of m basis functions $B(\cdot)$, it is expressed as

$$\mu_i = \sum_{j=1}^{m} a_j B_j(x_i), \quad i = 1, 2, \ldots, n.$$

Figure 1 shows the cubic basis functions for $m = 6$. For an description of basis functions, see Yoshimoto and Ichida (1973). Let

$$B = \begin{pmatrix} B_1(x_1) & \cdots & B_m(x_1) \\ \vdots & \ddots & \vdots \\ B_1(x_n) & \cdots & B_m(x_n) \end{pmatrix},$$

$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ and $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)'$. Then the model may be rewritten as

$$\boldsymbol{y} = B\boldsymbol{a}.$$

Therefore, in B-spline smoothing, we have to estimate $\boldsymbol{a}$ and σ^2.

When the parameters $\boldsymbol{a}$ and σ^2 are estimated using the maximum likelihood method, the more knots used, the more strongly dependent the estimated model is on the individual data pairs. Hence we estimate $\boldsymbol{a}$ and σ^2 by maximizing a penalized log-likelihood function

$$\begin{aligned} l_\lambda(\boldsymbol{a}, \sigma^2) &= \sum_{\alpha=1}^{n} \log f(y_i | x_i; \boldsymbol{a}, \sigma^2) - \frac{n\lambda}{2} \boldsymbol{a}' D_k' D_k \boldsymbol{a} \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\boldsymbol{y} - B\boldsymbol{a})'(\boldsymbol{y} - B\boldsymbol{a}) - \frac{n\lambda}{2} \boldsymbol{a}' D_k' D_k \boldsymbol{a}, \end{aligned}$$

which is taken into consideration for local variation in the log-likelihood function. In this equation, the smoothing parameter λ controls local variation in to estimated curve, and $\boldsymbol{a}' D_k' D_k \boldsymbol{a}$ is a penalty term related to the variation of a regression curve. D_k is the matrix representation of

kth differences,

$$D_k = \begin{pmatrix} (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k & 0 & \cdots & 0 \\ 0 & (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k \end{pmatrix},$$

where ${}_kC_i$ is the binomial coefficient. This penalized log-likelihood function cannot be solved via normal equations to estimate $\boldsymbol{a}$ and σ^2. Therefore, it is necessary to consider other maximization methods. We use the algorithm described Imoto and Konishi (1999a).

The algorithm for selecting the number of knots automatically is as follows.

Step1. Determine an arrangement and the initial number of knots.

Step2. Choose an initial value of λ, λ_0.

Step3. Search for the minimum value of the information criterion as a function of λ by applying the SPIDER algorithm.

Step4. Change the number of knots and seek the minimum value of the information criterion for each number of knots by iterating between Step2 and Step4.

Step5. Compare the information criteria minimal for each number of knots and carry out the smoothing using the number of knots that minimizes the criterion.

In this algorithm, we use the modified AIC (Eilers and Marix (1996)), which is defined by

$$\text{mAIC} = -2\sum_{i=1}^{n} \log f(y_i|x_i; \hat{\boldsymbol{a}}, \hat{\sigma}^2) + 2(\text{tr}S + 1),$$

where S is the hat matrix

$$S = B(B'B + n\hat{\sigma}^2\lambda D_k' D_k)^{-1}B',$$

and $\hat{\boldsymbol{a}}$ and $\hat{\sigma}^2$ are estimators of $\boldsymbol{a}$ and σ^2, respectively. The modified AIC is used because computation of its information is relatively simple. Moreover, we use the cubic basis functions and second differences as they are the most commonly employed.

3. Knot-placement strategies

3.1. Equidistant arrangement

The simplest method is to space the knots evenly, i.e. with equal width intervals between knots. We call this arrangement equidistant. Let $x_{max} = \max\{x_i \mid i = 1, 2, \ldots, n\}$ and $x_{min} = \min\{x_i \mid i = 1, 2, \ldots, n\}$. Then n_k knots are defined by

$$t(\alpha) = x_{min} + (\alpha - 4)(x_{max} - x_{min})/(\alpha - 7), \ \alpha = 1, 2, \ldots, n_k.$$

Note that data are not included outside $t(4)$ to $t(n_k - 3)$ in this arrangement. In actual use, x_{max} and x_{min} can be replaced by same other two values whose interval includes all data points. Moreover, the number of basis functions m is $n_k - 4$.

When the sample size n is large enough and the data are scattered uniformly, it is sensible to use this arrangement in B-spline smoothing. However, if the sample size n is small or the data are distributed nonuniformly on the x-axis, over fitting may occur in intervals with sparse data. We conjecture that the estimator of the coefficient $\boldsymbol{a}$ which is related basis function containing its interval is not obtained for a good one. Figures 2 and 3 display B-spline smoothing curves with the equidistant arrangement using different numbers of knots. Different curves result despite being fit to the same data. In these figures, $\circ$ and $+$ represent data and knots, respectively. Solid and broken lines illustrate the smoothing curve and fitted basis functions. In the case of Figure 2, $m = 14$, mAIC $= -108.667$ and $\lambda = 0.165658 \times 10^{-3}$. In Figure 3, $m = 19$, mAIC $= -121.7841$ and $\lambda = 0.35619 \times 10^{-5}$. Comparing the two information criteria, it seems that the smoothing curve in Figure 3 fits better than the one in Figure 2. But over fitting obviously occurs in the sparse interval in Figure 3. Figure 4 shows the information criterion as a function of the number of knots. Over fitting occurs frequency in the neighborhood of the end of points of x, but may also occur away from the end points, for instance, see Figure 5.

We posit that these situations can be avoided by shifting knots under the restriction of equidistant arrangement or by using a stopping rule, finishing if an interval does not include data, in the process of the optimization algorithm. However, to shift knots it is necessary to recalculate. Moreover, another problem is how to shift the knots. On the other hand, if the algorithm is stopped when an interval is found not to contain any data, then a good curve may not be obtained because the number of available knots is smaller than one without a stopping rule.

As stated above, there are limits the equidistant knot arrangement, so we must consider other arrangements.

3.2. Equipotent arrangement

In this subsection, we deal with a new knot-placement strategy, called the equipotent arrangement, where intervals between knots include the same number of data point as much as possible. These knots do not necessarily coincide with values of x_i. For that reason, the elements of B have value 0 on coinciding knots, unlike on non-coinciding ones. (Because, coinciding knots with data, several data are corresponding with the enp points of basis functions, hence, $B_j(x)$ is 0 in these data points.) Therefore, in the case of coinciding knots, if n is small then the estimation of $\boldsymbol{a}$ does not have a high degree of efficiency. Using this arrangement, we conjecture that a poor estimator of $\boldsymbol{a}$ is hard to be obtained than equidistant arrangement. Moreover, in the some case, to use the equipotent arrangement is not only an avoiding over fitting but also a getting more improvement for an information criterion than to use equidistant arrangement. However, since the interval width is reduced when the number of knots is large, the smoothing curve is extremely sensitive to the data in the narrow interval. Figure 6 depicts over fitting as with the previous situation. The results are $m = 19$, mAIC $= -131.765$ and $\lambda = 0.519465 \times 10^{-2}$. This smoothing curve is also constructed by knots whose number minimizes the information criterion. Examination of Figure 6 reveals that over fitting occurred in the narrow intervals.

3.3. Modified equipotent arrangement

In order to avoid over fitting as in the previous case, it is necessary that the interval widths do not become too narrow. Therefore, we consider another arrangement, called the modified equipotent arrangement. This arrangement is invoked if interval widths are smaller than a certain limit, in which case one of knots used to construct the interval is removed. Figure 7 shows a smoothing curve for same data in Figure 6 obtained by this method. Although 19 knots was optimal, only 15 were actually used ; four knots were removed because their intervals were smaller than a chosen limit $(x_{max} - x_{min})/20 = 1.162$. This number of knots is an optimized value which is obtained by using the information criterion. The resulting mAIC $= -120.286$ and $\lambda = 0.23070 \times 10^{-2}$. By comparing Figure 7 with Figures 5 and 6, it seems that using the modified equipotent arrangement avoids the over fitting that results from using the equipotent arrangement, but there is some loss in the information criterion.

At this time, we cannot put forth a method for specifying the limit on interval width. However, it seems from our experience that dividing the range of x by 20 provides a satisfactory value.

4. Monte Carlo simulation

In this section, we compare the three arrangements by iterating Monte Carlo simulations. Figures 8, 9 and 10 show the fluctuation in smoothing results with the equidistant, equipotent and modified equipotent arrangements. Normal random deviates were generated repeatedly. The values x_i $(i = 1, 2, \ldots, n)$ of the explanatory variable X were fixed at the observed values through out these experiments. Moreover, the optimal number of knots ranged from 8 to 20. We repeated the Monte Carlo simulation 10 times for each knot-placement strategy. From these figures, we can see that the equipotent and modified equipotent arrangements are avoided over fitting relative to the equidistant arrangement. Moreover, the modified equipotent arrangement produced smaller variation than the equipotent arrangement.

Next, we examined the fluctuation in smoothing curves quantitatively. In Figures 11 and 12, box plots illustrate the differences between smoothing curves and true trend. For each knot-placement strategy, we repeated the Monte Carlo simulation 500 times and calculated the mean squared errors between smoothing curves and true trend, dividing $x_{max} - x_{min}$ equally into 100 points. Figure 11 gives the case $n = 20$, and the case $n = 50$ is given in Figure 12. These figures show that the modified equipotent arrangement can avoid over fitting, though the average mean squared error is larger.

Acknowledgement

The authors wish to thank Dr. John Cologne, Radiation Effects Research Foundation, and Prof. Yasunori Fujikoshi, Hiroshima University, for his helpful comments.

References

[1] Atilgan, T. and Bozdogan, H. (1990). Selecting the number of knots in fitting cardinal B-splines for density estimation using AIC. *J.*

Japan Statist. Soc., 20, 179-190.

[2] De Boor, C. and Rice, J. R. (1968). Least squares cubic spline approximation I-fixed knots and II-fixed knots. *Purdue Univ. Reports*, CSD TR 20 and 21.

[3] Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89-121.

[4] Friedman, J. H. (1984). A variable span smoother. *Technical Report LCS5, Stanford University, Dept. of Statistics.*

[5] Ichida, K., Yoshimato, F. and Kiyono, T. (1976) Curve fitting by a piecewise cubic polynomial. *Computing*, **16**, 329-338.

[6] Imoto, S. and Konishi, S. (1999a). Nonlinear regression models using B-spline and information criteria. *Proceedings of the Institute of Statistical Mathematics* **47**, 359-373 (in Japanese).

[7] Imoto, S. and Konishi, S. (1999b). Estimation of B-spline nonlinear regression models using information criteria. *Japanese Journal of Applied Statistics*, **28**, 137-150 (in Japanese).

[8] Ohtaki, M. and Izumi, S. (1999). Globally convergent algorithm without derivatives for maximizing a multivariate function (in preparation).

[9] Yoshimoto, F. and Ichida, K. (1973). *Spline function and its application.* Kyouiku Syuppan (in Japanese).