Journal of Physics: Conference Series

PURPOSE-LED PUBLISHING™

PAPER • OPEN ACCESS

Generalized Cross Validation (GCV) in Smoothing Spline Nonparametric Regression Models

To cite this article: M Maharani and D R S Saputro 2021 *J. Phys.: Conf. Ser.* **1808** 012053

View the article online for updates and enhancements.

Generalized Cross Validation (GCV) in Smoothing Spline Nonparametric Regression Models

M Maharani[1*] and D R S Saputro[1]

[1]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Jl. Ir. Sutami 36A Kentingan, Jebres, Surakarta, 57126, Central Java, Indonesia.

*Email: mutiamaharani19@student.uns.ac.id

Abstract. A nonparametric regression model is utilized if the the regression curve does not contain information about the accepted shape and accepted curve is exist in a function. If any curve is given without the limitation of a certain functional form, a rough and non-unique curve will result. The smoothing spline can be utilized to remove rough curves in some segments by following a curve pattern. An approach that combines nonparametric regression and smoothing spline is known as the smoothing spline nonparametric regression model. The problem in estimating is the selection and determination of smoothing parameters obtained by taking into the sum of knots used and the position of the knots so that the Generalized Cross-Validation (GCV) method is required. A study was conducted on the smoothing spline nonparametric regression model on GCV. The method used in research is a literature study obtained from some articles, journals, and books that support research achievement. The results showed that with the GCV method the minimum GCV value was obtained which would determine how well the smoothing parameters shown by the estimator did not change significantly even though the number and position of the knots varied.

1. Introduction

The regression model is a show utilized to decide the numerical correlation between the predictor variable and the response variable. Assessing the regression curve gets to be an issue in regression analysis. In association with this estimate, there are two models utilized within the regression analysis that is the parametric and the nonparametric regression model. Parametric regression models are suitable in the event that the form of the regression curve is be discovered. The presumption of the parametric regression curve requires other sources accessible within the think about so that it can give point by point data. In case the form of the regression curve is not found or the instruction accessible around the regression curve is inadequately, at that point to estimate the regression curve it depends on the information so that a nonparametric regression model can be utilized. The use of constrained parametric regression yields inconsistent result. Within the nonparametric regression model, the curve is expected to fit in a function space where the function space choice is based on the nature of smoothness [1]. A process that can evacuate harsh curves by taking after a design is known as smoothing.

The development of smoothing techniques began in 1941 which was first introduced by Ezkeil. Smoothing aim is to minimize the diversity of data that does not affect so that the characteristics of the data will appear more clearly. Smoothing has gotten to be a common procedure in nonparametric methods utilized to assess function [2]. However, if any curves are given without any limitation on certain functional forms, the curve shape is inconsistent with the data. This is due to differences in the behavior of the function in each of its polynomial pieces. Spline technique is used to solve this problem by dividing the curve into several segments. Meanwhile, a model that combines smoothing and spline techniques is known as the smoothing spline model.

Smoothing spline is a nonparametric regression approach to obtain regression curve estimates [3]. Research conducted by [4] emphasized that estimation based on the smoothing spline technique has better results than kernel regression, where kernel regression is another nonparametric spline regression approaches. The main problem when estimating the smoothing spline regression function is selecting and determining the smoothing parameter [5]. According to [6] to obtain maximum results on smoothing parameters, the Generalized Cross Validation (GCV) method can be utilized. The GCV method is the superior of several methods that can be used to determine smoothing parameters because the calculation aspect is simpler and quite efficient [7]. In this study, was carried out for the GCV method in a nonparametric smoothing spline regression model.

2. Research Methods

The present research belongs to theory-based research which examines the smoothing parameters in the nonparametric smoothing spline regression model using the Generalized Cross Validation (GCV) method. The main procedure is to minimize the penalized least square model and determine the estimator function. This procedure includes estimating parameters, determining function estimators, and determining the matrix contained in the GCV method.

3. Result and Discussion

The study of spline regression was conducted by [8] using the spline technique. Based on the model formulated by [9] is to estimate the smoothing parameters to assess the smoothing spline regression curve. According to [10] there are various methods for selecting smoothing parameters and recommends GCV as a method for selecting spline smoothing parameters as computational effectiveness and precision of regression model functional coefficients. Based on the suggestion of [11], REML and GCV are a great smoothing parameter choice criteria for little and medium sample sizes. This procedure includes estimating parameters, determining function estimators, and determining the matrix contained in the GCV method.

3.1. Nonparametric regression

The model that is utilized when the regression curve is not found or does not follow a certain pattern is called a nonparametric regression model [12]. Nonparametric regression models are used to estimate regression curve that depend only on observed data. Suppose the predictor variable and response variable respectively (y_i, x_i), then the relationship between x_i and y_i is written as

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{3.1}$$

$f(x_i)$ is a regression function f to estimates y_i and ε_i is error that is expected to be normally distributed where variance σ^2 and zero mean.

Assuming a function with an unknown regression curve is the goal of nonparametric regression [7]. The regression curve is only expected to fit in a certain function space in the sense of smooth so that it has high flexibility [3]. There are several techniques for estimating the regression curve f in nonparametric regression, one of them is the smoohting spline.

3.2. Spline function on nonparametric regression

The spline is a segmented polynomial model where the segment properties provide better flexibility than the usual polynomial model. This property allows the spline regression model to fit properly to the local specifications of the data. The utilize of splines is centered on conduct or data patterns, which in certain sectors have different specifications from other sectors [12]. The m-order spline function with one explanatory variable is function that can generally be written in the form

$$f(x_i)=\beta_0+\sum_{j=1}^{m}\beta_j x^j+\sum_{j=1}^{k}\theta_j(x-K_j)_+^m, \qquad i=1,2,...,n. \qquad (3.2)$$

Then equation (3.2) is substituted to equation (3.1), the spline nonparametric regression equation is obtained which is written as

$$y_i=\beta_0+\sum_{j=1}^{m}\beta_j x^j+\sum_{j=1}^{k}\theta_j(x-K_j)_+^m+\varepsilon_i, \qquad i=1,2,...,n \qquad (3.3)$$

where β_0 is constant, β_j is the coefficient of the variable x_j, θ_j is the coefficient on the variable x_j cutting knots to$-k$, x^j is the independent variable of the order $-j$, K_j is the knot j on variable $x^j, j = 1,2,\dots,m$, k is the sum of knots and m is the order of the spline. In the matrix, equation (3.3) can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}=\begin{bmatrix} 1 & x_1 & \cdots & x_1^m & (x_1-K_1)_+^m & \cdots & (x_1-K_k)_+^m \\ 1 & x_2 & \cdots & x_2^m & (x_2-K_1)_+^m & \cdots & (x_2-K_k)_+^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^m & \cdots & x_n^m & (x_n-K_1)_+^m & \cdots & (x_n-K_k)_+^m \end{bmatrix}\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \\ \theta_1 \\ \vdots \\ \theta_k \end{bmatrix}+\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}. \qquad (3.4)$$

Equation (3.2) of the third term shows that the spline is a fragmented polynomial model (piecewise polynomial), with the spline function being continuous at the knots [13]. Knots is characterized as a central point within the spline function so that the curve shaped is fragmented at that point. The knot point is a point of intersection that shows changes in curve behaviour at different hoses [14].

Furthermore, if there is an error ε_i with mean and variance respectively $(0,\sigma^2)$ is assumed normal distribution, then y_i also same as mean and variance respectively $(f(x_i),\sigma^2)$. While the error of equation (3.4) can be expressed as

$$\boldsymbol{\varepsilon}=\mathbf{Y}-\mathbf{X}\boldsymbol{\beta} \qquad (3.5)$$

based on equation (3.5), the β parameter estimation is determined by the least square method by minimizing the number of squares error which is written as

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}=(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}). \qquad (3.6)$$

From equation (3.6), the estimation of the parameter β is obtained and expressed as

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

3.3. Smoothing spline nonparametric regression model

Smoothing spline is a function that can outline information well and has little error change [15]. The role of smoothing spline is to assess the regression function as a solution to the optimization problem obtained by minimizing the Penalized Least Square (PLS) function. Given $f(x_i)$ as a smooth function contained in a certain function space, in particular the Sobolev room $f \in W_2^v[a,b]$,

$$W_2^v[a,b]=\left\{f;\left(\int_a^b f^{(v)}(x)\right)^2 dt<\infty\right\}$$

where v is a positive number to solve the regression curve estimation and ε_i is error which is accepted to be normal distribution where mean and variance respectively $(0,\sigma^2)$ [1]. An optimal estimator when minimizing the penalized least square which is presented in the form

$$PLS=\sum_{i=1}^{n}(y_i-f(x_i))^2+\lambda\int_a^b(f''(x))^2dx \quad , \quad i=1,2,...,n \tag{3.7}$$

where $\sum_{i=1}^{n} \ \left(y_i-f(x_i)\right)^2$ is the sum of square error or a function of the range between the actual value and the estimated value, $\int_a^b \ \left(f''(x_i)\right)^2$ is a roughness penalty, which is a measure of the smoothness of the curve in mapping the data and λ is used to manage the adjust between the achievability of the data (the goodness of fit) and the continuity of the curve (penalty) known as smoothing parameter [16]. The λ has a very big influence on the PLS function [17]. The value λ change from zero up to $+\infty$, the arrangement changes from interpolation to linear models. When the value $\lambda \to +\infty$ Roughness penalty dominates the function PLS and the spline estimation is forced to be consistent. Meanwhile, the roughness penalty vanish from the function PLS, and the spline estimation interpolates the data when $\lambda \to 0$. Therefore, determining the smoothing parameter is very influential in assessing an unknown function.

The function estimator $\hat{f}$ can be determined by minimizing equation (3.7) which is written

$$\hat{f}=\underset{f\in W_2^m[a,b]}{Min}\left[\sum_{i=1}^{n}(y_i-f(x_i))^2+\lambda\int_a^b(f''(x_i))^2dx\right].$$

Used a much data n, with $Y=(Y_1,Y_2,\dots,Y_n)^T$ and f is a vector of $f(x_i)$, the equation (3.7) is determined in the form of a matrix which is written as

$$PLS=(\mathbf{Y}-\mathbf{f})^{\mathrm{T}}(\mathbf{Y}-\mathbf{f})+\lambda\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f} \tag{3.8}$$

where K is a penalty matrix having a specific structure defined as

$$\mathbf{K}=\mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^{\mathrm{T}} \tag{3.9}$$

Q is a matrix of sized $n\times(n-2)$ and R is matrix of sized $(n-2)\times(n-2)$. Matrix Q and R defined as

$$\mathbf{Q}=\begin{bmatrix} h_1^{-1} & 0 & \cdots & 0 \\ -h_1^{-1}-h_2^{-1} & h_2^{-1} & \cdots & 0 \\ h_2^{-1} & -h_2^{-1}-h_3^{-1} & \cdots & 0 \\ 0 & h_3^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_{n-1}^{-1} \end{bmatrix}_{n\times(n-2)} , h_i=x_{i+1}-x_i$$

$$\mathbf{R}=\begin{bmatrix} \frac{1}{3}(h_1+h_3) & \frac{1}{6}h_2 & \cdots & 0 \\ \frac{1}{6}h_2 & \frac{1}{3}(h_2+h_3) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{3}(h_{n-2}+h_{n-1}) \end{bmatrix}_{(n-2)\times(n-2)}$$

for further proof, see [18]. Next, the parameters estimator f in equation (3.8) can be obtained by minimizing the PLS function which is written as

$$\begin{aligned} PLS &= (\mathbf{Y}^{\mathrm{T}}-\mathbf{f}^{\mathrm{T}})(\mathbf{Y}-\mathbf{f})+\lambda\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f} \\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{Y}-\mathbf{Y}^{\mathrm{T}}\mathbf{f}-\mathbf{f}^{\mathrm{T}}\mathbf{Y}+\mathbf{f}^{\mathrm{T}}\mathbf{f}+\lambda\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f} \\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{Y}-2\mathbf{Y}\mathbf{f}^{\mathrm{T}}+\mathbf{f}^{\mathrm{T}}\mathbf{f}+\lambda\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f}. \end{aligned}$$

The PLS function becomes minimum if it is fulfilled $\frac{\partial PLS}{\partial f} = 0$, therefore

$$\frac{\partial PLS}{\partial \mathbf{f}} = 0$$

$$\mathbf{Y^TY - 2Yf^T + f^Tf + \lambda f^TKf} = 0$$

$$\mathbf{-2Y + 2f + 2\lambda Kf} = 0$$

$$\mathbf{-Y + f + \lambda Kf} = 0$$

$$\mathbf{f + \lambda Kf} = \mathbf{Y}$$

$$\mathbf{(I + \lambda K)f} = \mathbf{Y}$$

thus, the parameter estimate f is obtained and expressed as

$$\hat{\mathbf{f}} = \mathbf{(I + \lambda K)^{-1}Y} . \qquad (3.10)$$

Suppose $f = (f(x_1), \ldots, f(x_n))$ be a vector of the function value f at the points of knots $x_1, \ldots, x_n$ obtained by the smoothing spline estimator $\hat{f}_\lambda$ as an estimator of the function f or the estimated value for $Y = (y_1, \ldots, y_n)^T$and S_λ are known as smoothing matrix that are positive definite (symmetric), depend on the smoothing parameter but not depend on Y [18] which is written as

$$\hat{\mathbf{f}}_\lambda = \begin{bmatrix} \hat{f}_\lambda(x_1) \\ \hat{f}_\lambda(x_2) \\ \vdots \\ \hat{f}_\lambda(x_n) \end{bmatrix}_{(n\times 1)} = \mathbf{S}_\lambda \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(n\times 1)} , \hat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{Y} .$$

If $\hat{f}_\lambda$ is the estimator of the spline function and λ is the smoothing parameter in the spline regression, then selecting the optimal λ can be used the GCV method.

3.4. Generalized cross validation (GCV)

Determination of the optimal smoothing parameter is exceptionally critical to get a good curve estimator. To find out how well the estimator produced can be used the GCV method. In the smoothing spline nonparametric regression model, the method GCV_λ is written as

$$GCV_\lambda = n^{-1} \frac{\sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2}{(1 - n^{-1} trace[\mathbf{S}_\lambda])^2}$$
$$= \frac{MSE}{(n^{-1} trace[\mathbf{I} - \mathbf{S}_\lambda])^2} \qquad (3.11)$$

where S_λ is a matrix of size $n \times n$ which is defined as

$$\mathbf{S}_\lambda = \mathbf{(I + \lambda K)^{-1}}$$

where K is a matrix that is in accordance with equation (3.9), I as known as the identity matrix and $trace\ S_\lambda$ is the sum of the main diagonals of the expansion matrix S_λ of the smoothing parameter λ, while MSE it is the remaining average of square formulated as

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2 .$$

The minimum value of MSE indicates that the estimated value is close to the true value. It is assumed that the minimum value of MSE, MSE divided by the value $(n^{-1} trace[I - S_\lambda])^2$ with the value of S_λ influenced by λ, the optimum smoothing parameter is selected based on the small GCV value. Therefore, to obtain the optimal estimator of the spline function, it can be done by experimenting with the value of

λ $(0 < \lambda < 1)$ until the minimum GCV is obtained. The smoothing parameter in the spline regression will be optimum if the GCV_λ criterion value is relatively small, meaning that the estimator has not changed much.

4. Conclusion

Smoothing spline with the penalized least square is a combination of the smoothing function with the mean squares of the residues or MSE in the spline regression. The smoothing spline regression model greatly affects the value of the smoothing parameter. The smooth parameter value has an important role in deciding whether or not the resulting estimated regression curve is good. The method utilized in determining smoothing spline regression parameters is Generalized Cross Validation (GCV) which shows that with the GCV method minimum GCV value can be obtained which determines how well the optimum smoothing parameter marked by the estimator does not change significantly.

References

[1] Wahba G 1990 *Spline Models for Observational Data* (Philadelphia, Pa: Society for Industrial and Applied Mathematics)

[2] Adisantoso J 2010 Menentukan parameter pemulus pada model regresi smoothing spline *Media Staff Indones.*

[3] Eubank R L 1999 *Nonparametric Regression and Spline Smoothing* (New York: CRC Press)

[4] Aydin D 2007 A comparison of the nonparametric regression models using smoothing spline and kernel regression *World Acad. Sci. Eng. Technol.* **36** 253–7

[5] Cantoni E and Hastie T 2002 Degrees of freedom tests for smoothing splines *Biometrika* **89** 251–63

[6] Lee T C M 2003 Smoothing parameter selection for smoothing splines: a simulation study *Comput. Stat. Data Anal.* **42** 139–48

[7] Aydin D, Memmedli M and Omay R E 2013 Smoothing parameter selection for nonparametric regression using smoothing spline *Eur. J. Pure Appl. Math.* **6** 222–38

[8] Eubank R L 1988 *Spline Smoothing and Nonparametric Regression* (New York: M. Dekker)

[9] Krivobokova T, Crainiceanu C M and Kauermann G 2008 Fast adaptive penalized splines *J. Comput. Graph. Stat.* **17** 1–20

[10] Cao Y, Lin H, Wu T Z and Yu Y 2010 Penalized spline estimation for functional coefficient regression models *Comput. Stat. Data Anal.* **54** 891–905

[11] Aydın D and Memmedli M 2012 Optimum smoothing parameter selection for penalized least squares in form of linear mixed effect models *Optimization* **61** 459–76

[12] Hardle W 1990 *Applied Nonparametric Regression* (Cambridge: Cambridge University Press)

[13] Wang Y 2011 *Smoothing Splines: Methods and Applications* (New York: Chapman and Hall/CRC Press)

[14] Fan J and Yao Q 2008 *Nonlinear Time Series: Nonparametric and Parametric Methods* (New York: Springer Science and Business Media)

[15] Takezawa K 2005 *Introduction to Nonparametric Regression* (Hoboken, N.J: John Wiley and Sons)

[16] Lee T C M 2004 Improved smoothing spline regression by combining estimates of different smoothness *Stat. Probab. Lett.* **67** 133–40

[17] Ruppert D, Wand M P and Carroll R J 2003 *Semiparametric Regression* (Cambridge: Cambridge University Press)

[18] Green P J and Silverman Bernard W 1993 *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach* (New York: Chapman and Hall/CRC Press)