An estimating parameter of nonparametric regression model based on smoothing techniques

Autcha Araveeporn
Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
E-mail: kaautcha@hotmail.com

Abstract. This paper studies the estimating parameter of a nonparametric regression model that consists of the function of independent variables and observation of dependent variables. The smoothing spline, penalized spline, and B-spline methods in a class of smoothing techniques are considered for estimating the unknown parameter on nonparametric regression model. These methods use a smoothing parameter to control the smoothing performance on data set by using a cross-validation method. We also compare these methods by fitting a nonparametric regression model on simulation data and real data. The nonlinear model is a simulation data which is generated in two different models in terms of mathematical function based on statistical distribution. According to the results, the smoothing spline, the penalized spline, and the B-spline methods have a good performance to fit nonlinear data by considering the hypothesis testing of biased estimator. However the penalized spline method shows the minimum mean square errors on two models. As real data, we use the data from a light detection and ranging (LIDAR) experiment that contained the range distance travelled before the light as an independent variable and the logarithm of the ratio of received light from two laser sources as a dependent variable. From the mean square errors of fitting data, the penalized spline again shows the minimum values.

Keywords: B-spline, nonparametric regression, penalized spline, smoothing spline

1. Introduction

In statistical modelling, regression analysis is a statistical process for estimating parameters of the relationships between dependent and independent variables in terms of a regression function. However, regression analysis requires an assumption of the underlying regression function to be met. If an inappropriate assumption is used, it is possible to produce misleading results. To overcome this problem, the nonparametric regression is a choice to analyze data when the data are not meeting the assumption of regression analysis. The nonparametric regression is an alternative way for looking at scatter diagram smoothing to depict the relationship between dependent and independent variables. The single independent variable is called scatterplot smoothing it can be used to enhance the visual appearance to help our eyes pick out the trend in the plot.

The smoothing technique is a part of a method to estimate unknown parameters (trend or smoothing estimators) of nonparametric regression models. There are many popular smoothing techniques such as the smoothing spline [1,2], the penalized spline [3], and the B-spline [4]. The estimating parameters of these methods depend on the smoothing parameter which is controlled the trade-off between fidelity to the data and roughness of function. Smoothing Spline (SS) is a technique that estimates the natural polynomial spline by minimizing the penalized sum of squares based on a smoothing parameter. The penalized Spline (PS) smoother is approximated by minimizing the truncated power function on a low rank thin-plate spline depended on the smoothing parameter. The concept of

the B-spline is similar to the smoothing spline and penalized spline. This requires the piecewise constant B-spline that can be obtained from truncated counterparts by differencing the B-spline function.

In this paper, we consider the nonparametric regression model in Section 2, and use the smoothing spline, penalized spline, and B-spline methods to estimate the unknown parameter of nonparametric regression model in Section 3. In Sections 4 and 5, we show the estimation of these methods for simulation data and real data. The conclusion is presented in Section 6.

2. The nonparametric regression model

The nonparametric regression model consists of the cubic spline of piecewise polynomials function based on a function of independent variables ($S(x_t)$), error process (ε_t), and dependent variables (y_t) following

$$y_t = S(x_t) + \varepsilon_t, t = 1, 2, 3, ..., n. \tag{1}$$

The error process is assumed to follow the normal distribution with mean zero and variance one.

3. Method of smoothing techniques

The following smoothing techniques show the process to estimate parameters based on nonparametric regression model.

3.1. Smoothing spline method

Wahba [1] defined the natural polynomial spline $S(x_t) = S_K^m(x_t)$ as a real-valued function on $[a, b]$ with the aid of K so-called knots $-\infty \leqslant a < x_1 < x_2 < ... < x_K < b \leqslant \infty$. The class of m-order splines with domain $[a, b]$ will be denoted by $W^m[a, b]$.

The natural measure associated with the function $f \in W^m[a, b]$ that used to measure the roughness of curve which is called the quadratic penalty function given by

$$\int_b^a \left\{ S^{(m)}(x_t) \right\}^2 dx \tag{2}$$

where $S^{(m)}(x_t)$ is the mth derivative of $S(x)$ with respect to x.

Consider the simple nonparametric regression model, to estimate $\hat{S}(\cdot)$ minimizes $S_K^{(m)}(x_t)$ over the class of function $S(\cdot)$ following

$$S_K^{(m)}(x_t) = \min \sum_{t=1}^n \left\{ y_t - S(x_t) \right\}^2 + \lambda \int_a^b \left\{ S^{('')}(x_t) \right\}^2 dx \tag{3}$$

where $\lambda > 0$ denotes a smoothing parameter. In this study, we emphasize $m = 2$ so-called the natural cubic spline which is commonly considered in the statistical literature [2].

The natural cubic spline is given the value and second derivatives at each knots y_t as

$$\underset{\sim}{S} = S(x_t) = \beta_0 + \beta_1 S_1(x) + \ldots \beta_{n+3} S_{n+3}(x),$$

$$\underset{\sim}{\gamma} = S''(x_t), \quad t = 1, 2, 3, \ldots, n.$$

Let $\underset{\sim}{S}$ be the vector $(S_1, \ldots, S_{n+3})^{\mathrm{T}}$ and let $\underset{\sim}{\gamma}$ be the vector $(\gamma_1, \ldots, \gamma_{n+3})^{\mathrm{T}}$.

The condition of natural cubic spline depends on two matrices Q and R below

$$Q = \begin{pmatrix} h_1^{-1} & 0 & \cdots & 0 \\ -h_1^{-1}- & h_2^{-1} & \cdots & 0 \\ h_2^{-1} & & & \\ h_2^{-1} & -h_2^{-1}- & \cdots & 0 \\ & h_3^{-1} & & \\ 0 & h_3^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots h_{n-1}^{-1} \end{pmatrix}_{(n) \times (n-2),}$$

where $h_t = x_{t+1} - x_t$, for $t = 1, 2, \ldots, n$, then Q is a $n \times (n-2)$ matrix.

Matrix R is a symmetric $(n-2) \times (n-2)$ matrix with elements below

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

The matrix K can be decomposed by

$$K = QR^{-1}Q^{\mathrm{T}}. \tag{4}$$

The roughness penalty will satisfy

$$\int_a^b \left\{ S''(x_t) \right\}^2 dx_t = \gamma^{\mathrm{T}} R \gamma = S^{\mathrm{T}} K S$$

$$= \Omega_N(i, j). \tag{5}$$

To illustrate, it can be written in matrix form introduced by [2] as residual sum of squares (RSS)

$$RSS = \sum_{t=2}^{n} \{y_t - S(x_t)\}^2$$
$$= (y - S)^{\mathrm{T}}(y - S), \qquad (6)$$

where $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $S = (S(x_1), \ldots, S(x_n))^{\mathrm{T}}$. Letting N be a matrix with $N(i, j) = S_j(x_i)$ and $S = N\beta$.

The roughness penalty term $\int S''^2$ as Ω_N in Eq. (5) to obtain

$$\hat{S}_\lambda(x_t) = (y - N\beta)^{\mathrm{T}}(y - X\beta) + \lambda \beta^{\mathrm{T}} \Omega_N \beta \quad (7)$$

It therefore follows that Eq. (7) has a unique minimum, other smoothing spline estimator is obtained by

$$\hat{\beta} = (N^{\mathrm{T}}N + \lambda \Omega_N)^{-1} N^{\mathrm{T}} y,$$

then

$$\hat{S}_\lambda(x_t) = N(N^{\mathrm{T}}N + \lambda \Omega_N)^{-1} N^{\mathrm{T}}. \qquad (8)$$

In this paper, we also select the smoothing parameter using the method of generalized cross-validation (GCV) suggested by Wahba [5] and Craven and Wahba [6]. In practice, this step can be implemented by using the function of smooth.spline in the software R.

3.2. Penalized spline method

Eubank [7,8] introduced the regression spline that the local neighbourhoods are specified by a group of locations:

$$\tau_0, \tau_1, \tau_2, \ldots, \tau_K, \tau_{K+1} \qquad (9)$$

in the range of interval $[a, b]$, where $a = \tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} < b$. These locations are known as knots, and $\tau_r, r = 1, 2, \ldots, K$ are called interior knots.

A regression spline can be constructed using the k-th degree truncated power basis or called the B-spline basis with K knots $\tau_1, \tau_2, \ldots, \tau_K$:

$$1, x_t, \ldots, x_t^k, (x_t - \tau_1)_+^k, \ldots, (x_t - \tau_K)_+^k, \qquad (10)$$

where w_+^k denotes k-th power of the positive part of w where $w_+ = \max(0, w)$. The first $(k + 1)$ basis functions of the truncated power basis Eq. (12) are polynomials of degree up to k, and the others are all the truncated power functions of degree k. A regression spline can be expressed as

$$S(x_t) = \sum_{s=0}^{k} \beta_s x_t^s + \sum_{r=1}^{K} \beta_{k+r}(x_t - \tau_r)_+^k, \quad (11)$$

where $\beta_0, \beta_1, \ldots, \beta_{k+K}$ are the unknown coefficients to be estimated by a suitable loss minimization.

The penalized spline is a method to estimate a unknown smooth function using the truncated power function [9], and the penalized spline can be expressed as

$$S(x_t) = \sum_{j=0}^{m-1} \alpha_j x_s^j + \sum_{k=1}^{K} \beta_k (x_t - \tau_k)^{2m-1}, (12)$$

where $\beta = (\beta_1, \ldots, \beta_K)^{\mathrm{T}} \sim N\left(0, \sigma_\beta^2 \Omega^{-1/2}(\Omega^{1/2})^{\mathrm{T}}\right)$, and the (l, k) th entry of Ω is $|\tau_1 - \tau_k|^{2m-1}$ and only the coefficient of $|x_t - \tau_k|^{2m-1}$ are penalized so that a reasonably large order K can be used.

In this case, we focus $m = 2$, as the natural cubic spline, or called low-rank thin-plate spline which present of $S(\cdot)$ as

$$S(x_t) = \alpha_0 + \alpha_1 x_t + \sum_{k=1}^{K} \beta_k |x_t - \tau_k|^3, \qquad (13)$$

where $\theta = (\alpha_0, \alpha_1, \beta_1, \ldots, \beta_K)^{\mathrm{T}}$ is the vector of regression coefficients, and $\tau_1 < \tau_2 < \ldots < \tau_K$ are fixed knots. The number of knots, K can be selected using a cross-validation method or information theoretic methods (e.g., BIC or AIC).

This class of penalized spline smoothers $(\hat{S}(\cdot))$ may also be expressed as

$$\hat{S}(x_t) = C(C^{\mathrm{T}}C + \lambda^3 D)^{-1} C^{\mathrm{T}} y, \qquad (14)$$

where

$$C = \left[1 \; x_t \; |x_t - \tau_k|_{1 \leqslant k \leqslant K}^3\right]_{1 \leqslant t \leqslant n},$$

$$D = \begin{bmatrix} 0_{2 \times 2} & 0_{2 \times K} \\ 0_{K \times 2} & (\Omega_K^{1/2})^{\mathrm{T}} \Omega_K^{1/2} \end{bmatrix},$$

and $\lambda = \sigma_\beta^2 / \sigma_\varepsilon^2$ is a smoothing parameter. The penalized spline smoothers are estimated by using the Semi-Par package in the software R.

3.3. B-spline method

B-splines are very interesting as a basic function for univariate independent variable of nonparametric regression function. De Boor [10] gave an algorithm to compute B-spline of lower degree on piece wise polynomials function.

The mth degree of B-spline function are evaluated from $(m-1)^{\mathrm{th}}$ degree as

$$B_j^m(x) = \frac{x_t - \tau_j}{\tau_{j+m-1} - \tau_j} B_j^{m-1} + \frac{\tau_{j+m} - x_t}{\tau_{j+m} - \tau_{j+1}} B_{j+1}^{m-1}, \qquad (15)$$

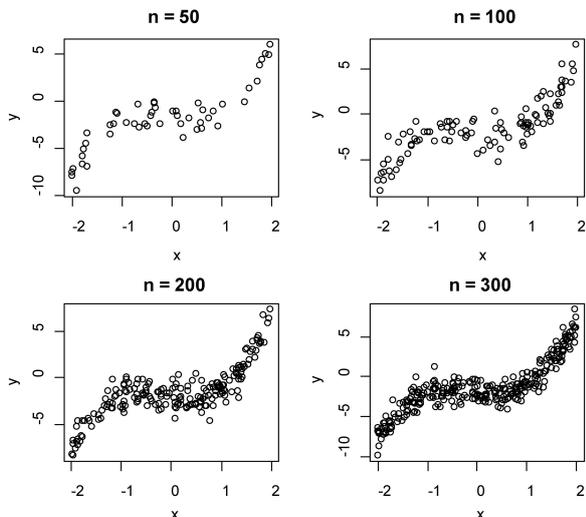Fig. 1. The scatter plot of dependent and independent variables on model 1.

Fig. 2. The scatter plot of dependent and independent variables on model 2.

where basis of order m with knots $\{B_i^m | i = M - m + 1, ..., M + K\}$, and auxiliary knots τ_j. B-splines base on non-zero over domain spanned by at most $M + 1$ knots. In this case, we focus the $m = 4$ or called the cubic B-spline with K knots has basis expansion as

$$S(x_t) = \sum_{j=1}^{K+4} B_j^4(x_t)\beta_j.$$

The nonparametric regression model can written in form of B-splines as

$$y_t = \sum_{j=1}^{K+4} B_j^4(x_t)\beta_j + \varepsilon_t, t = 1, 2, 3, ..., n. \quad (16)$$

In matrix form, B-splines can be written in form a linear model

$$B = \begin{bmatrix} B_1^4(x_1) \dots B_{K+4}^4(x_1) \\ \vdots \qquad\qquad \vdots \\ B_1^4(x_n) \dots B_{K+4}^4(x_n) \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

$$\text{and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The B-splines estimators are approximated by least square problems as

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_{K+4} \end{bmatrix} = (B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}y.$$
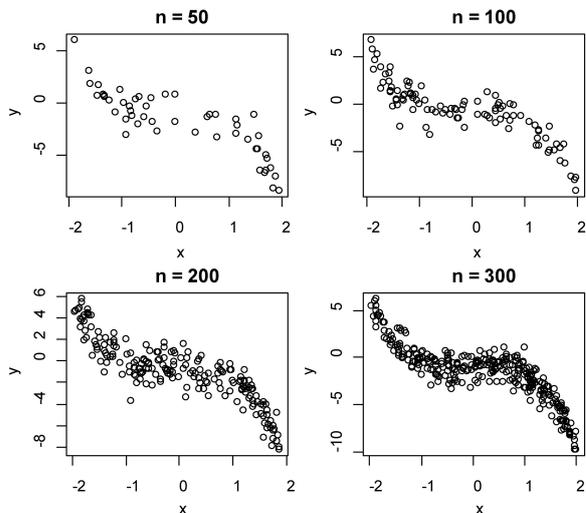
The B-spline and penalties are studied by Eilers and Marx [4] that advocate the use of the equally spaced knots, instead of the order statistics of the independent variable. The B-spline coefficients can be estimated as

$$\hat{\beta} = (B^{\mathrm{T}}B + \lambda D^{\mathrm{T}}D)^{-1}B^{\mathrm{T}}y, \quad (17)$$

where D is a banded matric which correspond to the difference penalty and denote by

$$D = \begin{bmatrix} 0 & 0 & 0\dots 0 & 0 \\ -1 & 1 & 0\dots 0 & 0 \\ 0 & -1 & 1\dots 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0\dots -1 & 1 \end{bmatrix}.$$

The fitting cubic B-splines are $\hat{S}(x_t) = \sum_{j=1}^{K+4} B_j^4 (x_t)\hat{\beta}_j$. The smoothing parameter λ choosing by minimizing the ordinary function or the generalized cross-validation function.

4. Simulation study

The nonlinear data of this study is simulated in two models for estimating the performance of smoothing techniques based on independent variables which are considered in the class of uniform distribution. These models in the process of construction a curve on mathematical function, that show the best fit to a series of data points. Figures 1 and 2 show the scatter plot of x_t and y_t on models 1 and 2 with 50, 100, 200, 300 sample sizes.

Table 1
The summary statistics of simulation studies with model 1 based on smoothing spline (SS), penalized spline (PS), and B-spline (BS)

Sample sizes	Methods	Mean	S.D.	LCI	UCI	t-statistic	p-values
$n = 50$	SS	-0.3344	6.5444	-0.9095	0.2405	-1.1428	0.2537
	PS	-0.3484	6.7454	-0.9411	0.2442	-1.1552	0.2468
	BS	0.3666	6.6935	-0.2214	0.9548	1.2249	0.2212
$n = 100$	SS	0.0719	4.2680	-0.3031	0.4469	0.3767	0.7065
	PS	0.0837	4.2029	-0.2855	0.4530	0.4454	0.6562
	BS	-0.0895	4.2152	-0.4599	0.2807	-0.4751	0.6349
$n = 200$	SS	-0.0954	2.8932	-0.3504	0.1595	-0.7353	0.4625
	PS	-0.0921	2.9123	-0.3480	0.1637	-0.7072	0.4798
	BS	0.0149	3.2685	-0.2299	0.2598	0.1200	0.9045
$n = 300$	SS	-0.1574	4.2673	-0.5347	0.2197	-0.8201	0.4125
	PS	-0.1441	4.1982	-0.5130	0.2247	-0.7676	0.4431
	BS	0.1475	1.8514	-0.2786	0.5738	0.6802	0.4966

Table 2
The summary statistics of simulation studies with model 2 based on smoothing spline (SS), penalized spline (PS), and B-spline (BS)

Sample sizes	Methods	Mean	S.D.	LCI	UCI	t-statistic	p-values
$n = 50$	SS	-0.0037	24.2710	-2.1363	2.1288	-0.0034	0.9972
	PS	-0.2582	23.8736	-2.3559	1.9893	-0.2419	0.8089
	BS	-0.3230	27.1906	-2.7121	2.0660	-0.2656	0.7906
$n = 100$	SS	-0.3966	15.8261	-1.7871	0.9939	-0.5603	0.5755
	PS	-0.2489	16.4557	-1.6947	1.1969	-0.3382	0.7353
	BS	0.5461	16.2381	-0.8806	1.9729	0.7520	0.4524
$n = 200$	SS	-0.3704	8.1031	-1.0838	0.3430	-1.0201	0.3082
	PS	-0.3447	7.8716	-1.0363	0.3468	-0.9792	0.3279
	BS	0.4035	8.5116	-0.3443	1.1514	1.0602	0.2896
$n = 300$	SS	0.0411	7.3081	-0.6055	0.6878	0.1250	0.9005
	PS	0.0408	7.0664	-0.5800	0.6617	0.1291	0.8973
	BS	0.0344	7.2205	-0.5999	0.6688	0.1066	0.9151

Model 1

$$S(x_t) = (x_t^3) - \cos(x_t) - \exp\left\{\frac{x_t}{1 + |x_t|}\right\},$$

$$x_t \sim Uniform(-2, 2), t = 1, 2, 3, ..., n$$

$$y_t = S(x_t) + \varepsilon_t, \varepsilon_t \sim Normal(0, 1),$$

$$t = 1, 2, 3, ..., n$$

Model 2

$$S(x_t) = \sin(x_t) - \exp\left\{\frac{x_t}{1 + |x_t|}\right\} - x_t^3,$$

$$x_t \sim Uniform(-2, 2), t = 1, 2, 3, ..., n$$

$$y_t = S(x_t) + \varepsilon_t, \varepsilon_t \sim Normal(0, 1),$$

$$t = 1, 2, 3, ..., n$$

The next step, the estimates of $S(x_t)$ or called $\hat{S}(x_t)$ are approximated from smoothing spline (SS), penalized spline (PS), and B-spline (BS) that used to compute the bias and MSE of $S(x_t)$ following

$$S_{bias} = \sum_{t=1}^{n} \left(\frac{\hat{S}(x_t) - S(x_t)}{S(x_t)}\right),$$

$$MSE = \frac{\sum_{t=1}^{n} \left[\hat{S}(x_t) - S(x_t)\right]^2}{n}.$$

The data are generated and repeated for fitting the model 500 times. A t-statistic is adopted to test that the mean of bias is equal the zero or called unbiased estimator. Tables 1 and 2 present the various summary statistics for the smoothing estimator obtained from three methods. The third and the fourth columns of these tables represent the sample mean and standard deviation of biases. The sample mean for the lower and upper bounds of the 95% confidence interval are given in the next two columns. The last two columns of these tables list the t-statistic, and p-values for hypothesis testing ($H_0 : \mu_{S_{bias}} = 0$ versus $H_0 : \mu_{S_{bias}} \neq 0$) that means when reject $H_0 : \mu_{S_{bias}} = 0$ the estimator SS, PS, and BS with bias. The histogram of the bias estimator of SS, PS, and BS in model 1 are presented in Figs 3–5, and model 2 are presented in Figs 6–8.

From Tables 1 and 2, by observing the p-values, the SS, PS, and BS provide asymptotically unbiased estimates for estimating parameter of $S(x_t)$ nearly for all sample sizes of two models. From the p-values for the two tables it is seen that are seen that the SS, PS, and
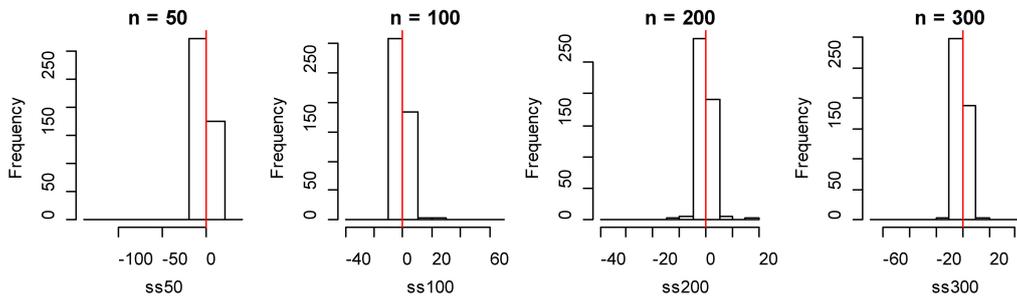
Fig. 3. Histogram of bias for fitting data of smoothing spline method with model 1.
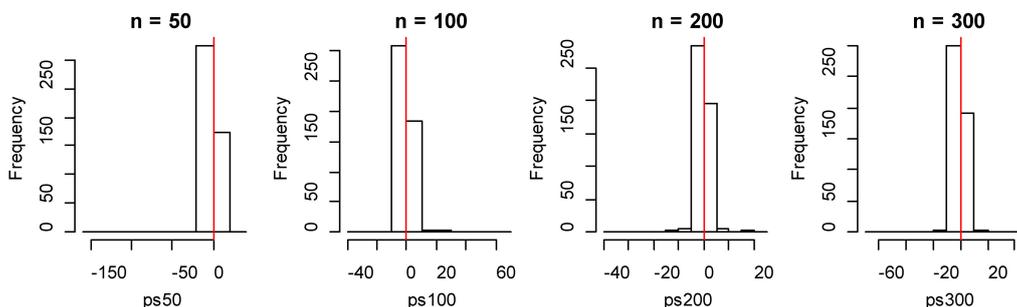

Fig. 4. Histogram of bias for fitting data of penalized spline method with model 1.
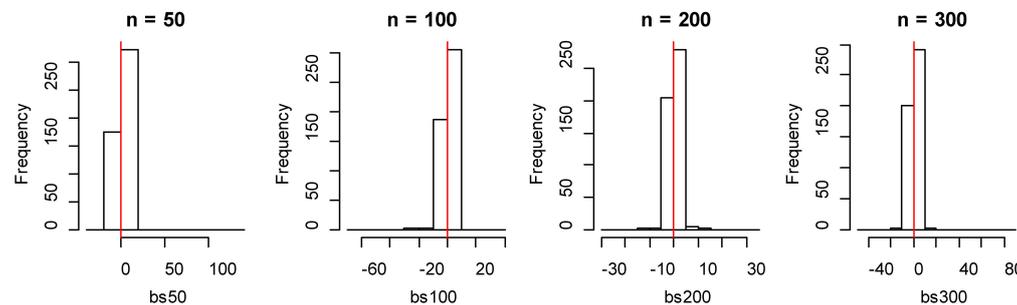

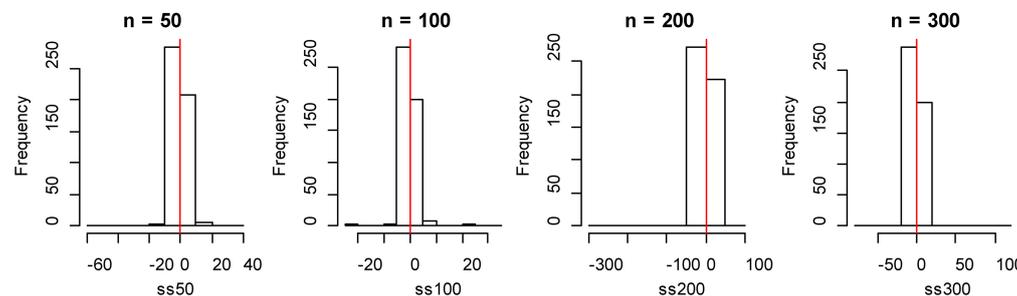Fig. 5. Histogram of bias for fitting data of B-spline method with model 1.

Fig. 6. Histogram of bias for fitting data of smoothing spline method with model 2.

BS of smoothing method have a good performance to fit data in a class of nonlinear data. From the histogram it is apparent that a standard deviation of relative bi-

ases increase with increasing sample sizes, so it makes the leptokurtic distribution. The average of MSE can answer the final question which smoothing method is
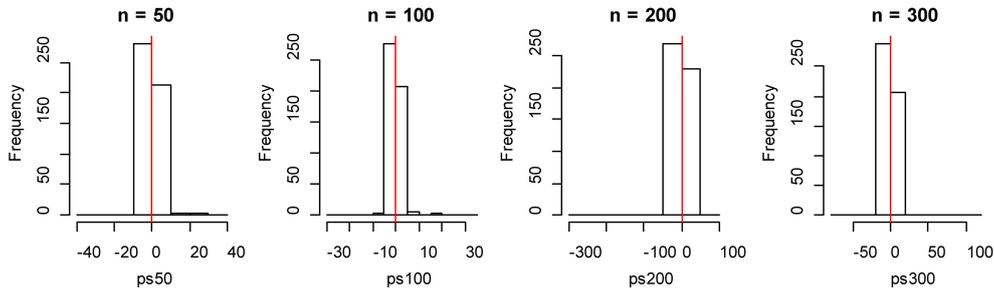
Fig. 7. Histogram of bias for fitting data of penalized spline method with model 2.
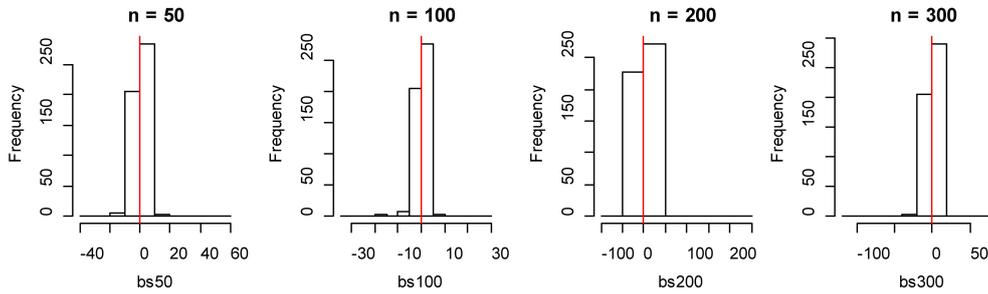

Fig. 8. Histogram of bias for fitting data of B-spline method with model 2.

Table 3
The average MSE of simulation studies with 3 models based on smoothing spline (SS), penalized spline (PS), and B-spline (BS)

Sample sizes	Methods	Model 1	Model2
$n = 50$	SS	0.8225	0.8348
	PS	**0.7712**	**0.7846**
	BS	0.9174	0.9364
$n = 100$	SS	0.9087	0.9029
	PS	**0.8810**	**0.8719**
	BS	0.9632	0.9508
$n = 200$	SS	0.9528	0.9521
	PS	**0.9387**	**0.9439**
	BS	0.9784	0.9840
$n = 300$	SS	0.9666	0.9648
	PS	**0.9640**	**0.9615**
	BS	0.9899	0.9877

Fig. 9. The plot of LIDAR data frame and model fitting of SS, PS, and BS methods.

the best estimator. Table 3 shows the average MSE for fitting 500 times on two models, and it can be seen that the PS method shows the minimum of average MSE for all sample sizes and models.

5. Application of real data

In this section, we consider the application of smoothing method based on SS, PS, and BS methods that we developed in the previous section. As the real data, we use the data frame which consists of 221 observations from a light detection and ranging (LIDAR) experiment. This data frame contains the range dis-
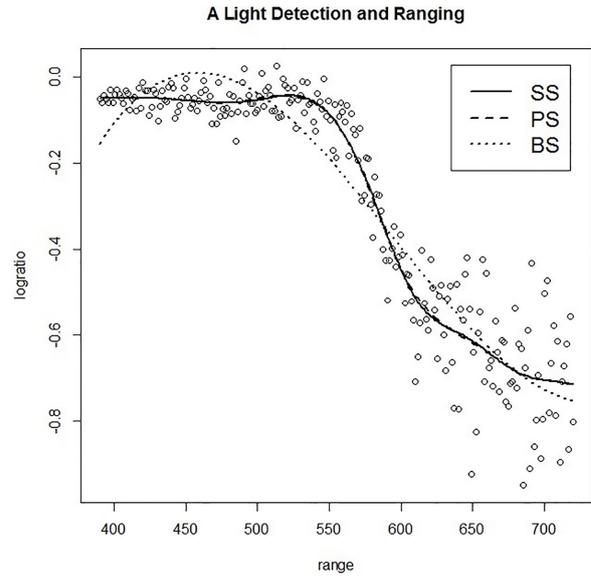
tance travelled before the light is reflected back to its source and logarithm of the ratio of received light from two laser sources as shown in the plot in Fig. 9.

After fitting the model, the estimating values play on a plot of light detection of ranging. It can be seen that the SS and PS interpolate in mass data more than the

BS method that followed the MSE values such as SS = 0.006016, PS = 0.006010, and BS = 0.009288. The minimum of MSE is the PS which is closed the SS as the result on Table 3.

6. Conclusion

In this section, we used the smoothing techniques of SS, PS, and BS methods based on nonparametric regression models. Through a Monte Carlo simulation study, we evaluated the smoothing estimator of SS, PS, and BS methods. For hypothesis testing based on the p-value, the fitting values supported the null value, and showed that the smoothing estimators work reasonably well for all methods, but the PS shows the minimum of average MSE.

Acknowledgments

References

[1] Wahba G. Spline Models for Observational Data, SIAM: Philadelphia; 1990.

[2] Green PJ, Silverman BW. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, Chapman and Hall: London; 1994.

[3] Ruppert D, Wand MP, Carroll RJ. Semiparametric Regression, Cambridge University Press: New Your; 2003.

[4] Eilers PHC, Marx BD. Flexible Smoothing with B-splies and Penalties, Statistical Science 11(2) (1996), 89–102.

[5] Wahba G. A survey of some smoothing problems and the method of generalized cross-validation for solving them, In Proceeding of the Conference on the Application of Statistic, 1976, pp. 507–523.

[6] Craven P, Wahba G. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, Numerische Mathematik 31 (1979), 377–403.

[7] Eubank RL. Spline Smoothing and Nonparametric Regression, Marcel Dekker: New York; 1988.

[8] Eubank RL. Nonparametric Regression and Spline Smoothing, Marcel Dekker: New York; 1999.

[9] Ruppert D, Carroll RJ. Spatial-adaptive penalties for spline fitting Australian and New Zealand Journal of Statistics 42 (2000), 205–224.

[10] De Boor C. A Practical Guide to Splines, Springer: Berlin; 1978.