ELSEVIER

Statistics & Probability Letters 25 (1995) 105–111

STATISTICS & PROBABILITY LETTERS

Behavior near zero of the distribution of GCV smoothing parameter estimates

Grace Wahba*, Yuedong Wang

Department of Statistics, University of Wisconsin – Madison, 1210 West Dayton street, Madison, WI 53706-1685, USA

Received December 1993; revised September 1994

Abstract

It has been noticed by several authors that there is a small but non-zero probability that the GCV estimate $\hat{\lambda}$ of the smoothing parameter in spline and related smoothing problems will be extremely small, leading to gross undersmoothing. We obtain an upper bound to the probability that the GCV function, whose minimizer provides $\hat{\lambda}$, has a (possibly local) minimum at 0. This upper bound goes to 0 exponentially fast as the sample size gets large. For the medium- to small-sample case we study this probability both by Monte Carlo evaluation of a formula for the exact probability that the GCV function has a minimum at 0 as well as by replicated calculations of $\hat{\lambda}$.

Keywords: Smoothing spline; GCV; Smoothing parameter

1. Introduction

Consider the model

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1,\ldots,n, \ t_i \in [0,1], \tag{1}$$

where $\epsilon = (\varepsilon_1,\ldots,\varepsilon_n)^T \sim N(0,\sigma^2 I_{n\times n})$, σ^2 unknown and $f \in W_m$, where $W_m = \{f : f, f', \ldots, f^{(m-1)}$ absolutely continuous, $\int_0^1 (f^{(m)})^2 \, dt < \infty\}$. The smoothing spline $\hat{f}_\lambda$ is the minimizer in W_m of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int_0^1 (f^{(m)}(t))^2 \, dt. \tag{2}$$

As is well known, the smoothing parameter λ controls the trade-off between the goodness of fit and the roughness, and it is important to choose an appropriate λ. The GCV estimate $\hat{\lambda}$ of λ is the minimizer of the

* Corresponding author. Tel.: (608)262-3620/2598. Fax: (608)262-0032. E-mail: Wahba@stat.wisc.edu. Supported by NSF Grant DMS9121003 and NEI Grant R01 EY09946.

SSDI 0167-7152(94)00211-8

GCV function $V(\lambda)$ given by

$$V(\lambda) = \frac{1}{n}\|(I - A(\lambda))y\|^2 \Big/ \left[\frac{1}{n}\mathrm{tr}(I - A(\lambda))\right]^2, \tag{3}$$

where $A(\lambda)$ is the so-called 'hat' matrix which relates $\hat{g}_\lambda = (\hat{f}_\lambda(t_1),\ldots,\hat{f}_\lambda(t_n))^T$ to $y = (y_1,\ldots,y_n)^T$ by $\hat{g}_\lambda = A(\lambda)y$. See [Wahba (1990)] and the references cited there. Several authors (for example [Wahba, 1983]; [Thompson et al., 1989]) have noted that, with small sample sizes, there is a usually small, but positive probability that GCV (erroneously) chooses an 'extremely small' $\hat{\lambda}$. These extremely small cases can be easily distinguished from the bulk of the other cases by the fact that estimates of the variance based on them are several orders of magnitude below the true σ^2 while the other cases give reasonable estimates of σ^2. Theoretical results on the optimality of $\hat{\lambda}$ when λ is estimated by GCV have apparently needed to assume that $\hat{\lambda}$ is the minimum of $V(\lambda)$ in $[\lambda_n, \infty]$, where $\lambda_n \to 0$ but not too fast, see [Cox (1988)], [Nychka (1988)] and [Nychka (1990)]. [Wahba (1990)] conjectured that the theoretical distribution of $\hat{\lambda}$ has a small mass point at $\lambda = 0$ for moderate n which decreases with n. We think that the remainder of the distribution of $\hat{\lambda}$ has, for practical purposes, a more or less concentrated density about an optimum λ, as contrasted to a 'very fat tail' on the left. The remarks in Thompson et al. tend to agree with this, although they have apparently plotted their density based on samples of $\hat{\lambda}$ with a tail rather than a mass point. Since both the numerator and the denominator of $V(\lambda)$ behave like λ^2 as $\lambda \to 0$, special care is needed to search for the minimum at or near 0. Thus, the question of whether the distribution of $\hat{\lambda}$ consists of a mass point at 0 plus a concentrated density, as opposed to a density with a very long left tail, is not straightforward to settle numerically.[1] It is known that there is a non-zero probability that $V(\lambda)$ has a (possibly local) minimum at 0. The main purpose of this note is to provide theoretical and empirical results on the decrease of the size of the probability of this local minimum at 0 as n becomes large. We also observed that the distribution of $\hat{\lambda}$ may be described as having a mass point at 0 plus a fairly concentrated distribution away from 0 in the examples studied.

In Section 2, we obtain an asymptotic approximation to the probability p that $V(\lambda)$ has a local minimum at 0, to use as an (asymptotic) upper bound for the probability that GCV chooses $\hat{\lambda} = 0$. We prove that this probability goes to zero at least exponentially fast. This asymptotic result is independent of the unknown function and the variance of the noise. [Nychka (1991)] discusses the behavior of an asymptotic distribution for $\hat{\lambda}$ which is suggestive that any mass at 0 disappears quickly with n. In Section 3, we carry out a computational approximation to the exact probability that $V(\lambda)$ has a local minimum at 0. This probability does depend on the unknown f and σ^2 as well as n. We carry out the calculations for several combinations of f and σ^2 and $n = 32, 64$ and 128. This calculation is easy to do for any hypothesized f, σ^2 and design points. We then use these same f's, σ^2 and n to replicate data and obtain samples of $\hat{\lambda}$ by a search. The empirical fraction of extremely small $\hat{\lambda}$ roughly agree with the computed probabilities of a local minimum at 0, and, furthermore, there are no 'moderately small' cases. The computational approximation and the results of the search give results which agree and for which the mass point at 0 shrinks quickly as n becomes large.

2. The probability p

We use the same notation as in [Wahba (1990)]. Let $\phi_\nu(t) = t^{\nu-1}/(\nu-1)!$, $\nu = 1,\ldots,m$; $R^1(s,t) = \int_0^1 (s-u)_+^{m-1}(t-u)_+^{m-1}\,du/[(m-1)!]^2$; $T_{n\times m} = \{\phi_\nu(t_i)\}_{i=1\ \nu=1}^{n\ \ m}$; $\Sigma_{n\times n} = \{R^1(t_i,t_j)\}_{i=1\ j=1}^{n\ \ n}$. Let the QR decomposition of T be

$$T = (Q_1 : Q_2)\begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{4}$$

[1] Thompson et al. also have a few multiple minima in their examples, which had $n = 47$.

where Q_1 is $n \times m$ and Q_2 is $n \times (n-m)$, $Q = (Q_1 : Q_2)$ is orthogonal and R is upper triangular. Let the eigenvector eigenvalue decomposition of $Q_2'\Sigma Q_2$ be UDU', where U is $(n-m) \times (n-m)$ orthogonal and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n-m})$. Let $\Gamma = Q_2 U$ and $z = (z_1, \ldots, z_{n-m})^{\mathrm{T}} = \Gamma^{\mathrm{T}} y$. We can write[2]

$$V(\lambda) = \frac{1}{n}\sum_{\nu=1}^{n-m}\left(\frac{n\lambda z_\nu}{\lambda_\nu + n\lambda}\right)^2 \Bigg/ \left(\frac{1}{n}\sum_{\nu=1}^{n-m}\frac{n\lambda}{\lambda_\nu + n\lambda}\right)^2.$$

Following Thompson et al., we have

$$V'(0) = 2n^2\left[\sum_{\nu=1}^{n-m}\frac{1}{\lambda_\nu^2}\sum_{\mu=1}^{n-m}\frac{z_\mu^2}{\lambda_\mu^2} - \sum_{\nu=1}^{n-m}\frac{1}{\lambda_\nu}\sum_{\mu=1}^{n-m}\frac{z_\mu^2}{\lambda_\mu^3}\right] \Bigg/ \left[\sum_{\nu=1}^{n-m}\frac{1}{\lambda_\nu + n\lambda}\right]^3,$$

and so

$$p = P(V'(0) > 0) = P\left(\sum_{\nu=1}^{n-m}\frac{1}{\lambda_\nu^2}\sum_{\mu=1}^{n-m}\frac{z_\mu^2}{\lambda_\mu^2} > \sum_{\nu=1}^{n-m}\frac{1}{\lambda_\nu}\sum_{\mu=1}^{n-m}\frac{z_\mu^2}{\lambda_\mu^3}\right). \tag{5}$$

In this section we investigate the large sample property of p. We only consider the special case f periodic, $\hat{f}_\lambda$ a periodic spline and $t_i = i/n$. This case is simple to analyze while still giving insight into the general case.[3] Then, $\Gamma^{\mathrm{T}} = W$, the scaled discrete Fourier transform matrix. W is listed in Table 4.1 on p. 54 of [Wahba (1990)]. Let $\boldsymbol{g} = (f(1/n), \ldots, f(n/n))^{\mathrm{T}}$, $\boldsymbol{h} = W\boldsymbol{g}$ and $\boldsymbol{e} = W\boldsymbol{\epsilon}$. Then $\boldsymbol{e} \sim \mathrm{N}(0, \sigma^2 I_{n\times n})$. Let f_ν's be the Fourier coefficients of f. Then $f_\nu = \int_0^1 f(t)\cos 2\pi\nu t\,\mathrm{d}t = [\sum_{\mu=1}^{n}(1/n)f(\mu/n)\cos 2\pi\nu(\mu/n)](1 + \mathrm{o}(1))$. Similar approximations hold for the sin's. So we have $\boldsymbol{f} = (f_1, \ldots, f_n)^{\mathrm{T}} = (1/\sqrt{n})W\boldsymbol{g}(1 + \mathrm{o}(1))$. Substituting in the definition of $\boldsymbol{h}$ we get $\boldsymbol{h} = \sqrt{n}\boldsymbol{f}(1 + \mathrm{o}(1))$. Following the arguments in Wahba (1990, Chs. 2 and 5), we have that $\lambda_\nu = n(2\pi\nu)^{-2m}(1 + \mathrm{o}(1))$. Note for later reference that if qm is an integer, we have $\int_0^1 (f^{(qm)}(t))^2\mathrm{d}t = \sum(2\pi\nu)^{2mq} f_\nu^2$, where, with some abuse of notation, the sum is taken over both the sine and cosine coefficients. It is easy to prove that

$$\sum_{\nu=1}^{n}\frac{1}{\lambda_\nu^k} = \frac{(2\pi)^{2km}}{n^k}\sum_{\nu=1}^{n}\nu^{2km} = \frac{(2\pi)^{2km}}{(2km+1)}n^{2km+1-k}(1 + \mathrm{o}(1)).$$

Lemma 1. For any integer l, $\sum_{\nu=1}^{n}\nu^l e_\nu^2/\sigma^2 n^{l+1}/(l+1) \xrightarrow{\mathrm{p}} 1$ as $n \to \infty$.

Proof. The moment generating function of the l.h.s. is

$$E\exp\left\{t\sum_{\nu=1}^{n}\nu^l e_\nu^2/\sigma^2\frac{n^{l+1}}{l+1}\right\} = \exp\left\{-\frac{1}{2}\sum_{\nu=1}^{n}\ln\left(1 - 2t\nu^l\Big/\frac{n^{l+1}}{l+1}\right)\right\}$$
$$= \exp\left\{t\sum_{\nu=1}^{n}\nu^l\Big/\frac{n^{l+1}}{l+1} + \mathrm{o}(1)\right\} \to \exp\{t\}, \quad n \to \infty.$$

Let

$$A_n = \frac{1}{\sigma^2 c}\sum_{\nu=1}^{n}\left(\frac{\nu^{6m}}{2m+1} - \frac{n^{2m}\nu^{4m}}{4m+1}\right)e_\nu^2,$$

[2] This kind of formula applies to the general penalized least-squares problem, see Wahba (1990).

[3] In this case $I - A(0)$ is of rank $n-1$ rather than $n-m$ so that the sums in $V'(0)$ go from 1 to $n-1$. With some abuse of notation we will just write 1 to n.

where

$$c = \sum_{v=1}^{n} \left(\frac{v^{6m}}{2m+1} - \frac{n^{2m} v^{4m}}{4m+1} \right) = \frac{4m^2}{(2m+1)(4m+1)^2(6m+1)} n^{6m+1}(1 + o(1)).$$

The moment generating function of A_n is

$$\begin{aligned} & E \exp\{tA_n\} \\ &= \exp\left\{ -\frac{1}{2} \sum_{v=1}^{n} \ln\left[1 - \frac{2t}{c} \left(\frac{v^{6m}}{2m+1} - \frac{n^{2m} v^{4m}}{4m+1} \right) \right] \right\} \\ &= \exp\left\{ \frac{1}{2} \sum_{v=1}^{n} \left[\frac{2t}{c} \left(\frac{v^{6m}}{2m+1} - \frac{n^{2m} v^{4m}}{4m+1} \right) + \frac{1}{2} \frac{4t^2}{c^2} \left(\frac{v^{6m}}{2m+1} - \frac{n^{2m} v^{4m}}{4m+1} \right)^2 \right] + o\left(\frac{1}{n} \right) \right\} \\ &= \exp\left\{ t + \frac{t^2}{2n} \kappa^2 + o\left(\frac{1}{n} \right) \right\}, \end{aligned}$$

where

$$\kappa^2 = \frac{(4m+1)^2(6m+1)^2(56m^2+6m+1)}{2m^2(8m+1)(10m+1)(12m+1)}.$$

So $\sqrt{n}(A_n - 1)/\kappa \xrightarrow{d} Z$ as $n \to \infty$, where Z is the standard normal random variable.

Let us go back to the original problem. We have

$$\begin{aligned} p &= P\left(\sum_{v=1}^{n} \frac{1}{\lambda_v^2} \sum_{\mu=1}^{n} \frac{z_\mu^2}{\lambda_\mu^2} > \sum_{v=1}^{n} \frac{1}{\lambda_v} \sum_{\mu=1}^{n} \frac{z_\mu^2}{\lambda_\mu^3} \right) \\ &= P\left(\frac{n^{2m}}{4m+1} \left[\sum_{v=1}^{n} v^{4m} e_v^2 + \sum_{v=1}^{n} v^{4m} h_v^2 + 2 \sum_{v=1}^{n} v^{4m} h_v e_v \right] > \right. \\ &\quad \left. \frac{1}{2m+1} \left[\sum_{v=1}^{n} v^{6m} e_v^2 + \sum_{v=1}^{n} v^{6m} h_v^2 + 2 \sum_{v=1}^{n} v^{6m} h_v e_v \right] \right) (1 + o(1)) \\ &= P(A_n < B_n)(1 + o(1)), \end{aligned} \tag{6}$$

where

$$B_n = \frac{1}{\sigma^2 c} \left[\frac{1}{2m+1} \left(\sum_{v=1}^{n} v^{6m} h_v^2 + 2 \sum_{v=1}^{n} v^{6m} h_v e_v \right) - \frac{n^{2m}}{4m+1} \left(\sum_{v=1}^{n} v^{4m} h_v^2 + 2 \sum_{v=1}^{n} v^{4m} h_v e_v \right) \right].$$

Suppose that $|h_v| \leqslant b\sqrt{n}(2\pi v)^{-q}(1+o(1))$, where b is a constant independent of v and n. For f to be a periodic function with $\int_0^1 (f^{(m)}(t))^2 \, dt \leq \infty$, we need $\int_0^1 (f^{(m)}(t))^2 dt = \sum (2\pi v)^{2m} f_v^2 < \infty$, that is, $q > \frac{1}{2} + m$. It is easy to prove that if $q > \frac{3}{4}$, $\sum_{v=1}^{n} v^{6m} h_v^2 \sim n^{6m+2-2q} = n^{-1/2} o(n^{6m+1})$, and $\mathrm{Var}(\sum_{v=1}^{n} v^{6m} h_v e_v) \sim n^{12m+2-2q} = n^{-1} o(n^{12m+2})$. So $\sum_{v=1}^{n} v^{6m} h_v e_v = n^{-1/2} o_p(n^{6m+1})$, where for any random variables X_n and Y_n, $X_n > 0$, we define $Y_n = o_p(X_n)$ if $Y_n / X_n \to 0$ in probability as $n \to \infty$. Similarly, if $q > \frac{3}{4}$, $n^{2m} \sum_{v=1}^{n} v^{4m} h_v^2 \sim n^{6m+2-2q} =$

$n^{-1/2}o(n^{6m+1})$ and $n^{2m}\sum_{v=1}^{n} v^{4m}h_v e_v = n^{-1/2}o_p(n^{6m+1})$. Combining them together, we have $n^{1/2}B_n = o_p(1)$. From (6),

$$\begin{aligned} p &= P(\sqrt{n}(A_n - 1) < \sqrt{n}(B_n - 1))(1 + o(1)) \\ &= P(\sqrt{n}(A_n - 1) < -\sqrt{n})(1 + o(1)) \\ &= \Phi\left(-\frac{\sqrt{n}}{\kappa}\right)(1 + o(1)), \end{aligned}$$

where Φ is the standard normal distribution. Note that $\sum_{v=1}^{n} v^{4m} f_v^2 = \int_0^1 (f^{(2m)}(t))^2 dt(1 + o(1))$ and $\sum_{v=1}^{n} v^{6m} f_v^2 = \int_0^1 (f^{(3m)}(t))^2\, dt(1 + o(1))$, provided that these converge. So the above approximations are obvious if $\int_0^1 (f^{(3m)}(t))^2\, dt < \infty$. But in the approximations above, we only need that $q > \frac{3}{4}$. So the result holds provided only that $\sum_{v=1}^{\infty} v^{2mq} f_v^2 \leqslant \infty$ for some $q > 3/4$. When n is large, a simple upper bound for p is $(2\kappa/\sqrt{2\pi n})\exp\{-n/2\kappa\}$. So p goes to zero at least exponentially fast. □

3. Simulations

Notice that the bound we get in Section 1 does not depend on σ and f. It is not a sharp bound for p. It is a good approximation when n is large since the dropped term B_n is asymptotically much smaller than the dominating term A_n provided $q > \frac{3}{4}$. But it may not be a good approximation for modest n. A local minimum of V at zero does not imply that GCV chooses $\hat{\lambda} = 0$. On the other hand, if GCV numerically picks a very small $\hat{\lambda}$ (not zero), that does not imply that V has a local minimum at zero. $V(\lambda)$ may have a local minimum near zero. It would be nice to know how much difference there is between these two possibilities.

The value p in (5) can be calculated by Monte Carlo methods if we know the true function f and variance σ^2. We can simply generate observations y from the true function. Then we can calculate z_v's and λ_v's. Comparing the summations inside (5), we record 1 if the inequality holds and 0 otherwise. Repeat this process N times. The Monte Carlo estimate of p is simply the frequency of 1's.

In the following, we conduct simulations to calculate Monte Carlo estimates of p and compare them to the number of times that GCV chooses a very small λ.

The experimental design is the same as in Wahba (1983). Three functions are used:

$$\text{Case 1:} \quad f(t) = \frac{1}{3}\beta_{10,5}(t) + \tfrac{1}{3}\beta_{7,7}(t) + \tfrac{1}{3}\beta_{5,10}(t);$$

$$\text{Case 2:} \quad f(t) = \tfrac{6}{10}\beta_{30,17}(t) + \tfrac{4}{10}\beta_{3,11}(t);$$

$$\text{Case 3:} \quad f(t) = \tfrac{1}{3}\beta_{20,5}(t) + \tfrac{1}{3}\beta_{12,12}(t) + \tfrac{1}{3}\beta_{7,30}(t);$$

where

$$\beta_{p,q}(t) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} t^{p-1}(1-t)^{q-1}, \quad 0 \leq t \leq 1.$$

Cases 1, 2 and 3 have 1, 2 and 3 bumps, respectively. They reflect increasing difficulty to fit with a spline. The experiment consists of $3 \times 3 \times 5 = 45$ combinations of Case 1, 2, 3, n=32, 64, 128 and σ=0.0125, 0.025, 0.05, 0.1 and 0.2. In all cases, $t_i = i/n$. Data are generated for 100 replications of each of these 45 combinations. In all simulations in this report, random numbers are generated using the Fortran routine rnor of the Core Mathematics Library (Cmlib) from the National Bureau of Standards. Spline fits were calculated using RKPACK,[4] see [Gu (1989)].

[4] The RKPACK default golden section search was used to find the minimizer of $V(\lambda)$.

Table 1
Number of replications out of 100 total that have $\hat{\lambda}$ smaller than -14 in $\log_{10}$ scale

	$n = 128$			$n = 64$			$n = 32$		
	1	2	3	1	2	3	1	2	3
$\sigma = 0.0125$	0	0	0	0	4	4	8	70	100
$\sigma = 0.025$	0	0	0	0	3	4	7	24	57
$\sigma = 0.05$	0	0	0	0	3	4	7	12	22
$\sigma = 0.1$	0	0	0	0	0	2	7	8	11
$\sigma = 0.2$	0	0	0	0	0	0	7	7	8

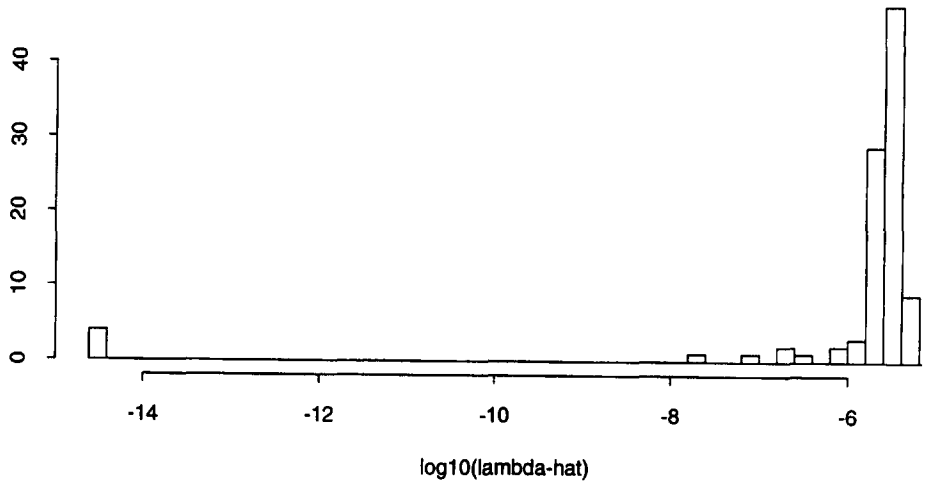

Fig. 1. Histogram of 100 samples of $\hat{\lambda}$ for the $n = 64, \sigma = 0.05$, Case 3 example of Table 1, illustrating the gap between 'extremely small' and the rest of the distribution.

Table 1 lists the number of replications out of 100 simulation replications that have a GCV estimate of $\hat{\lambda}$ smaller than -14 in $\log_{10}$ scale. GCV estimates of $\hat{\lambda}$ for all other cases are bigger than -9 in $\log_{10}$ scale. All cases in Table 1 also have ratio $\hat{\sigma}/\sigma < 0.001$, where $\hat{\sigma}^2 = ||I - A(\hat{\lambda})y||^2/\mathrm{tr}(I - A(\hat{\lambda}))$. All other cases had $\hat{\sigma}/\sigma > 0.1$. So in practice, it is easy to identify these 'extremely small' cases if σ is known to within an order of magnitude. Fig. 1 gives a histogram of the 100 replicates of $\hat{\lambda}$ for the $n = 64, \sigma = 0.05$, Case 3 example of Table 1, which shows clearly the gap between the 'extremely small' cases and the other cases.

The Monte Carlo estimates of p's in percentages are listed in Table 2. For these calculations we took $N = 10^6$, with the same seed for each table entry. The similarity between the two tables suggests that there is little gap between the probability that $V(\lambda)$ has a local minimum at zero and the probability that GCV chooses an extremely small $\hat{\lambda}$. In other words, the results suggest that they happen at the same time.

Tables 1 and 2 also indicate that for small sample sizes, the probability that GCV chooses a very small $\hat{\lambda}$ and the probability that $V(\lambda)$ has a local minimum at zero do depend heavily on σ, n and the shape of the function. The probabilities decrease as sample size increases, σ increases or the number of bumps decreases. The upper bound obtained in Section 1 does not work for the sample size $n = 32$.

Table 2
Monte Carlo estimates of p's in percentages

	$n = 128$			$n = 64$			$n = 32$		
	1	2	3	1	2	3	1	2	3
$\sigma = 0.0125$	0.3	0.3	0.3	3.0	3.5	3.3	11.5	67.9	98.9
$\sigma = 0.025$	0.3	0.3	0.3	3.0	3.1	3.1	10.8	25.6	61.5
$\sigma = 0.05$	0.3	0.3	0.3	3.0	3.0	3.0	10.7	13.9	23.3
$\sigma = 0.1$	0.3	0.3	0.3	3.0	3.0	3.0	10.6	11.4	13.5
$\sigma = 0.2$	0.3	0.3	0.3	3.0	3.0	3.0	10.6	10.8	11.3

Acknowledgements

This research was supported by the National Science Foundation under Grant DMS-9121003 and the National Eye Institute under Grant R01 EY09946.

References

Cox, D.D. (1988), Approximation of method of regularization estimatior, *Ann. Statist.*, pp. 694–712.

Gu, C. (1989), RKPACK and its applications: fitting smoothing spline models, Technical Report 857, Dept. of Statistics, Univ. of Wisconsin – Madison.

Nychka, D. (1988), Bayesian confidence intervals for smoothing splines, *J. Amer. Statist. Assoc.* **83**, 1134–1143.

Nychka, D. (1990), The average posterior variance of a smoothing spline and a consistent estimate of the average squared error, *Ann. Statist.* **18**, 415–428.

Nychka, D. (1991), Choosing a range for the amount of smoothing in nonparametric regression, *J. Amer. Statist. Assoc.* **86**, 653–664.

Thompson, A.M., J.W. Kay and D.M. Titterington (1989), A cautionary note about crossvalidatory choice, *J. Statist. Comput. Simulation* **33**, 199–216.

Wahba, G. (1983), Bayesian confidence intervals for the cross validated smoothing spline, **45**, 133–150.

Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59 (SIAM, Philadelphia).