Sample Size Calculations for
Smoothing Splines

Based on Bayesian Confidence Intervals

Yuedong Wang*

October 17, 1997

Abstract

Bayesian confidence intervals of a smoothing spline are often used to distinguish two
curves. In this paper, we provide an asymptotic formula for sample size calculations
based on Bayesian confidence intervals. Approximations and simulations on special
functions indicate that this asymptotic formula is reasonably accurate.

Key Words: Bayesian confidence intervals; sample size; smoothing spline.

*Address: Department of Statistics and Applied Probability, University of California, Santa Barbara, CA
93106-3110. Tel.: (805)893-4870. Fax: (805)893-2334. E-mail: yuedong@pstat.ucsb.edu. Supported by the
National Institute of Health under Grants R01 EY09946, P60 DK20572 and P30 HD18258.
1 Smoothing Splines and Their Bayesian Confidence Intervals

Consider the model

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \cdots, n, \quad x_i \in [0, 1],$$  \hspace{1cm} (1)

where $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T \sim N(0, \sigma^2 I_{n \times n})$, $\sigma^2$ unknown. Assume $f \in W_2^m$, where

$$W_2^m = \{ f : f^{(v)} \text{ absolutely continuous}, \ v = 0, \cdots, m-1, \ \int_0^1 (f^{(m)}(x))^2 dx < \infty \}. \hspace{1cm} (2)$$

A smoothing spline $\hat{f}_\lambda$ is the minimizer of

$$\min_{f \in W_2^m} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_0^1 (f^{(m)}(x))^2 dx \right\}. \hspace{1cm} (3)$$

When $\lambda$ is fixed, $\hat{f}_\lambda = (\hat{f}_\lambda(x_1), \cdots, \hat{f}_\lambda(x_n))^T$ is a linear function of $y = (y_1, \cdots, y_n)^T$: $\hat{f}_\lambda = A(\lambda) y$, where $A(\lambda)$ is the so called “hat”, or influence, matrix.

Wahba (1978) showed that $\hat{f}_\lambda$ equals to the posterior mean of the following Bayes model. Suppose that $f$ in (3) is a sample path from the Gaussian process

$$F_\xi(x) = \sum_{j=1}^m \frac{\tau_j x^{j-1}}{(j-1)!} + b^\frac{1}{2} \int_0^x \frac{(x - s)^{m-1}}{(m-1)!} dW(s), \hspace{1cm} (4)$$
where $W(\cdot)$ is a standard Weiner process and $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_m)^T \sim N(0, \xi I_{m \times m})$. Wahba (1978) showed that with $n\lambda = \sigma^2/b$,

$$
\hat{f}_{\lambda}(x) = \lim_{\xi \to \infty} E(F_{\xi}(x)|y), \quad \sigma^2A(\lambda) = \lim_{\xi \to \infty} \text{Cov}(F_{\xi}|y),
$$

(5)

where $F_{\xi} = (F_{\xi}(x_1), \cdots, F_{\xi}(x_n))^T$.

This connection between a smoothing spline and the posterior mean and variance led Wahba (1983) to propose the $(1 - \alpha)100\%$ Bayesian confidence intervals for $\{f(x_i)\}_{i=1,n}$ as

$$
\hat{f}_{\lambda}(x_i) \pm z_{\alpha/2}\sqrt{\hat{\sigma}^2 a_{i,i}}, \quad i = 1, \cdots, n,
$$

(6)

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard Normal distribution. $\hat{\lambda}$ is an estimate of $\lambda$, which can be selected by data-based procedures such as GML, GCV and other methods (see Wahba (1990)). $\hat{\sigma}^2 = \|I - A(\hat{\lambda})y\|^2/tr(I - A(\hat{\lambda}))$ is an estimate of $\sigma^2$. Both simulations (Wahba 1983, Wang and Wahba 1995) and theory (Nychka 1988, Nychka 1990) suggested that these Bayesian confidence intervals have good frequentist properties for $f \in W_2^n$ provided $\hat{\lambda}$ is a good estimate of the $\lambda$ which minimizes the predictive mean squared error. The intervals must be interpreted “across the function”, rather than pointwise. More precisely, assume $f \in W_2^n$, instead of a realization of the stochastic process (4). Let $\tau_n$ be a point
randomly selected from \( \{x_i\}_{i=1}^{n} \). The average coverage probability has the property

\[
\frac{1}{n} \sum_{i=1}^{n} P\left( |f(x_i) - \hat{f}_\lambda(x_i)| \leq z_{\alpha/2} \sqrt{\sigma^2 \hat{a}_{i}} \right) \approx P\left( |f(\tau_n) - \hat{f}_\lambda(\tau_n)| \leq z_{\alpha/2} \hat{\sigma} \sqrt{tr(A(\lambda))/n} \right) \approx 1 - \alpha. \tag{7}
\]

See Nychka (1988) for details.

Bayesian confidence intervals are often used to distinguish two curves for various reasons; for example, to decide whether a spline estimate is more suitable than a particular parametric regression estimate. A parametric regression model may be considered unsuitable if its estimate is outside of a large portion of the confidence intervals of a smoothing spline estimate. In section 2, we give an asymptotic formula for sample size calculations based on these Bayesian confidence intervals. In section 3, we calculate sample sizes for a special case and conduct simulations to evaluate the performance of these approximations.

## 2 Sample Size Needed to Distinguish Two Curves

Suppose \( x_i \)'s are iid random samples from a density function \( w \). Assume \( w \) is strictly positive on \([0, 1]\) and \( w \in C^\infty[0, 1] \). Suppose we have two different curves \( f \) and \( g \) defined on \([0, 1]\), with some data generated from curve \( f \) plus Gaussian noise. That is, the data have the form (1). We can construct Bayesian confidence intervals for \( f \) as in (6). Under the following assumptions (see Nychka (1990) for discussions about these assumptions):

**Assumption 1**: \( \hat{\lambda} \) is the minimizer of the GCV function over the interval \([\lambda_n, \infty)\), where \( \lambda_n \sim n^{-2m/5} \log(n)^m \);
**Assumption 2:** $f$ is such that for some $\gamma > 0$, \[ \frac{1}{n} \sum_{i=1}^{n} (Ef_{i}(x_{i}) - f(x_{i}))^2 = \gamma \lambda^2 (1 + o(1)) \] uniformly for $\lambda \in [\lambda_n, \infty)$; Nychka (1990) proved that $\hat{\sigma}^2 \rightarrow_p \sigma^2$ and $trA(\lambda)/n \rightarrow 0$ as $n \rightarrow \infty$. Thus the average Bayesian interval length approaches to zero. More and more points of $g$ lie outside of the confidence intervals as the sample size increases. The question is: what is the smallest sample size such that if we randomly choose a point $\kappa$ according to the density function $w$, the probability of $g(\kappa)$ is outside the 100(1 $- \alpha$)% Bayesian confidence interval of $f(\kappa)$ is larger than $1 - \beta$? More precisely, what sample size is needed for

\[
P(g(\kappa) \notin C(\alpha, f(\kappa))) \geq 1 - \beta,
\]  

where $C(\alpha, f(\kappa))$ is the $1 - \alpha$ Bayesian confidence interval of $f$ at $\kappa$. $\alpha$ and $1 - \beta$ correspond to the size and power in the usual hypothesis test.

Define random variable $d(f, g, \kappa) = |f(\kappa) - g(\kappa)|$. Denote by $D$ the cumulative distribution function of $d(f, g, \kappa)$ and let $d_\beta$ be the $\beta$ percentile of $D$. Suppose that $d_\beta > 0$. As argued in Nychka (1988), $a_{i,i} \approx trA(\lambda)/n, i = 1, \cdots, n$. Then

\[
C(\alpha, f(\kappa)) \approx (\hat{f}_\lambda(\kappa) - z_{\alpha/2} \sigma \sqrt{trA(\lambda)/n}, \hat{f}_\lambda(\kappa) + z_{\alpha/2} \sigma \sqrt{trA(\lambda)/n}).
\]  

If $z_{\alpha/2} \sigma \sqrt{trA(\lambda)/n} \leq d_\beta$, we have

\[
P(g(\kappa) \notin C(\alpha, f(\kappa)))
\]
\[ \approx P(|\hat{f}_\lambda(\kappa) - g(\kappa)| > z_{\alpha/2}\sigma\sqrt{\text{tr}A(\hat{\lambda})/n}) \]
\[ \approx P(|f(\kappa) - g(\kappa)| > z_{\alpha/2}\sigma\sqrt{\text{tr}A(\hat{\lambda})/n}) \]
\[ \geq P(|f(\kappa) - g(\kappa)| > d_\beta) \]
\[ = 1 - \beta. \]

Therefore the sample size we need is
\[ n \approx \left( \frac{z_{\alpha/2}\sigma}{d_\beta} \right)^2 \text{tr}A(\hat{\lambda}). \tag{10} \]

The quantity \( \text{tr}A(\hat{\lambda}) \) is sometimes called the degrees of freedom for signal. It goes to infinity as \( n \) goes to infinity if \( f \notin \mathcal{H}_0 \). Suppose Assumptions 1 and 2 hold. From Lemma 3.1 and Lemma 3.2 of Nychka (1990), as \( n \to \infty \), we have
\[ \text{tr}A(\lambda) = a l_1 \lambda^{-\frac{1}{2m}} (1 + o(1)), \quad \text{uniformly for } \lambda \in [\lambda_n, \infty], \tag{11} \]
\[ \hat{\lambda} = \left( \frac{a l_2 \sigma^2}{n \gamma 4m} \right)^{\frac{2m}{4m}} (1 + o_p(1)), \tag{12} \]

where \( a = \int_0^1 (w(x))^{\frac{1}{2m}} dx / \pi \)

\[ ^1 \text{Please note the definition of } \alpha \text{ in Nychka (1990) should be } \alpha = \int_0^1 (g(v))^{1/2m} dv / \pi. \]

\[ \int_0^1 (g(v))^{1/2m} dv / \pi. \]
Under some conditions (cf. Wahba (1983)), \( \gamma = \int_0^1 (f^{(2m)}(x))^2 \text{d}x \). Suppose that \( C = \sup_{x \in [0,1]} |f(x)| > 0 \). Let \( \tilde{f}(x) = f(x)/C, \tilde{g}(x) = g(x)/C, \tilde{d}_\beta \) be the \( \beta \) percentile of the distribution of \( d(\tilde{f}, \tilde{g}, \kappa) \). Then \( \tilde{d}_\beta = d_\beta/C \) and \( \tilde{\gamma} = \int_0^1 (\tilde{f}^{(2m)}(x))^2 \text{d}x = \gamma/C^2 \). Replacing \( d_\beta \) and \( \gamma \) in (13), we have

\[
 n \approx B \tilde{\gamma}^{1/2m} \left( \frac{z_{\alpha/2}}{\tilde{d}_\beta} \right)^{4m+1 \over 2m} \left( \frac{\sigma}{C} \right)^2 , \tag{14}
\]

where

\[
 B = a \frac{4m+1}{4m} l_1^{-1 \over 2m} (4m)^{1 \over m} \tag{15}
\]

is a constant. From (14), we can see that the sample size increases with the decrease of the size \( \alpha \), with the increase of the power \( 1 - \beta \), with the increase of the roughness \( \tilde{\gamma} \), and with the decrease of the signal to noise ratio \( C/\sigma \).

\( d_\beta \) could be calculated or estimated through Monte Carlo simulation if \( f, g \) and \( w \) are known. In practice, we do not know \( f, g, w \) and \( \sigma \). A pilot study or prior knowledge are necessary in order to estimate sample sizes adequately. One should also take into account the possibility of near interpolation if GCV method will be used to estimate the smoothing parameter (Wahba and Wang (1993)).
3 Simulations

In this section, for particular functions of $f$ and $g$, we calculate sample sizes based on formula (14). We check these asymptotic results by an approximation to $trA(\lambda)$ and simulations.

Let

$$f(x) = C \sin 2\pi x, \quad g(x) = 0. \quad (16)$$

Let $w(x) = 1$, that is, $\kappa$ is uniformly distributed on $[0,1]$. We have $a = 1/\pi$. Suppose we want to fit a cubic spline ($m=2$). We have $l_1 = \int_0^\infty dx/(1+x^4) \approx 1.11$, $l_2 = \int_0^\infty dx/(1+x^4)^2 \approx 0.83$, $B = l_1^{1/8} l_2^{-1/8} 8^{1/8}/\pi \approx 0.48$, $\tilde{\gamma} = \int_0^1 (\sin^{(2)} 2\pi x)^2 dx = 2^7 \pi^8$, $\tilde{d}_\beta = \sin(\beta \pi /2)$. Then

$$n \approx B 2^\frac{7}{2} \pi \left( \frac{z_{\alpha/2}}{\sin(\beta \pi/2)} \right)^2 \left( \frac{\sigma}{C} \right)^2. \quad (17)$$

Sample sizes based on (17) for $\alpha = 0.9, 0.95, 0.99$, $1-\beta = 0.75, 0.9, 0.95$ and $C/\sigma = 5, 2, 1$ are listed in Table 1 under the columns of “Asymp.” All numbers in the table are the smallest integers that greater than the right hand side of (17).

When $n$ is even and design points are equally spaced, we can approximate $trA(\lambda)$ by

$$trA(\lambda) \approx 1 + 2 \sum_{\nu=1}^{n/2-1} \frac{1}{1 + \lambda (2\pi \nu)^{2m}} + \frac{1}{1 + \lambda (\pi n)^{2m}}. \quad (18)$$

See Wahba (1983) for details about this approximation. For comparison, we calculate the sample sizes as the smallest even number such that $z_{\alpha/2} \sqrt{trA(\lambda)/n} \leq d_\beta$ and list them
in table 1 under columns “Approx.” Again we replace $\lambda$ by $\hat{\lambda}$. We can see that these two methods give almost the same sample sizes.

Put Table 1 here

Next, we conduct simulations to evaluate how good are these approximations to the real sample sizes needed. We only consider some interesting cases: $\alpha = 0.9, 0.95$, $1 - \beta = 0.9$, $C = 1$, and $\sigma = 0.5, 1$. For a particular case, the number of sample size $n$ based on (17) is used. For example, $n = 137$ for $\alpha = 0.9$, $1 - \beta = 0.9$ and $\sigma = 0.5$. We first generate an iid sample $x_i$’s of size $n$ from uniform distribution on $[0, 1]$. Observations $y_i$’s are generated by $y_i = \sin 2\pi x_i + \epsilon_i$, where $\epsilon_i$’s are iid samples from $N(0, \sigma^2)$. We fit $\{(y_i, x_i)\}_{i=1}^n$ by a cubic spline with the smoothing parameter $\lambda$ estimated by the GML method. Spline fits are calculated using RKPACK (Gu 1989). We then construct Bayesian confidence intervals for the smoothing spline estimate. The power is estimated by the proportion on $[0, 1]$ that zero is outside these Bayesian confidence intervals. Repeat this process 100 times, we have 100 power estimates. Box plots of these powers of all four cases are shown in Figure 1. All simulations have the estimated powers near the nominal values (dotted lines). The medians of these powers are smaller than the nominal values. One possible reason is that the true “across the function” coverage of Bayesian confidence intervals exceed their nominal values.

Put Figure 1 here
4 Acknowledgements

The author is grateful to Grace Wahba for her constant encouragement and Douglas Bates and Brian Yandell for reading this part in Wang’s thesis. The author thanks Douglas Nychka for helpful discussions and a referee for comments that greatly improved the paper.
References


Table 1: Sample sizes for different $\alpha$’s, $\beta$’s and signal to noise ratios.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.9</td>
<td>.75</td>
<td>3</td>
<td>2</td>
<td>19</td>
<td>20</td>
<td>73</td>
<td>74</td>
</tr>
<tr>
<td>.9</td>
<td>.9</td>
<td>22</td>
<td>24</td>
<td>137</td>
<td>138</td>
<td>546</td>
<td>546</td>
</tr>
<tr>
<td>.9</td>
<td>.95</td>
<td>104</td>
<td>104</td>
<td>644</td>
<td>646</td>
<td>2576</td>
<td>2578</td>
</tr>
<tr>
<td>.95</td>
<td>.75</td>
<td>5</td>
<td>2</td>
<td>28</td>
<td>28</td>
<td>109</td>
<td>110</td>
</tr>
<tr>
<td>.95</td>
<td>.9</td>
<td>33</td>
<td>34</td>
<td>203</td>
<td>204</td>
<td>809</td>
<td>810</td>
</tr>
<tr>
<td>.95</td>
<td>.95</td>
<td>153</td>
<td>154</td>
<td>956</td>
<td>956</td>
<td>3822</td>
<td>3824</td>
</tr>
<tr>
<td>.99</td>
<td>.75</td>
<td>8</td>
<td>8</td>
<td>50</td>
<td>50</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>.99</td>
<td>.9</td>
<td>60</td>
<td>60</td>
<td>374</td>
<td>376</td>
<td>1496</td>
<td>1498</td>
</tr>
<tr>
<td>.99</td>
<td>.95</td>
<td>283</td>
<td>284</td>
<td>1766</td>
<td>1768</td>
<td>7064</td>
<td>7068</td>
</tr>
</tbody>
</table>
Figure 1: $\alpha = 0.1$, $1 - \beta = 0.9$. Left: estimates of the powers. Right: estimates of the coverages of the Bayesian confidence intervals. Dotted lines are nominal values.