Smoothing Parameter Selection for Smoothing Splines:

A Simulation Study*

Thomas C. M. Lee†

June 12, 2001

Abstract

Smoothing splines are a popular method for performing nonparametric regression. Most important in the implementation of this method is the choice of the smoothing parameter. This article provides a simulation study of several smoothing parameter selection methods, including two so-called “second generation” methods. To the best of the author’s knowledge, the empirical performances of these two second generation methods have never been reported in the literature. Empirical conclusions from and recommendations based on the simulation results will be provided.

Keywords: exact double smoothing, nonparametric regression, plug-in methods, risk estimation, roughness penalty, smoothing parameter, smoothing splines

*Running Title: Smoothing Spline Smoothing Parameter Selection
†Postal: Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA; Email: tlee@stat.colostate.edu; Phone: (970) 491 2185; Fax: (970) 491 7895.
1 Introduction

This article studies the problem of nonparametric regression using smoothing splines. Suppose observed are \( n \) pairs of measurements \((x_i, y_i), i = 1, \ldots, n\), relating to the model

\[ y_i = f(x_i) + \epsilon_i, \quad a < x_1 < \ldots < x_n < b, \quad \epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2), \]

where \( f(x) \) is an unknown function of interest. A smoothing spline estimate \( \hat{f}_\lambda \) for \( f \) is defined as the minimizer of the penalized criterion

\[
\frac{1}{n} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(x)\}^2 \, dx.
\]

In the above \( \lambda \) is a positive constant known as the smoothing parameter. It controls the trade-off between the bias and the variance of \( \hat{f}_\lambda \). For general references on smoothing splines, consult, for examples, Eubank (1988), Green & Silverman (1994) and Wahba (1990).

It is widely known that \( \lambda \) has a crucial effect on the quality of \( \hat{f}_\lambda \). Therefore obtaining an appropriate choice of \( \lambda \) is an important problem. The main purpose of this article is to, via numerical experiments, evaluate and compare the finite sample performances of several data-dependent methods for choosing \( \lambda \). Of course it is impossible to exhaust all possible experimental settings, but we shall use the approach adopted by Wand (2000) to alleviate this problem. The idea is to change one experimental factor (e.g., noise level) at a time so that patterns can be more easily detected.

Altogether six smoothing parameter selection methods will be compared. Three of them are "classical" while the remaining three are so-called "second generation". The three classical methods are cross-validation (CV), generalized cross-validation (GCV) and Mallows'
$C_p$ criterion. In both the local linear regression and kernel density estimation contexts, it has been known that these classical selectors tend to be highly variable and also have the tendency to undersmooth. In order to repair these drawbacks, various second generation methods, such as plug-in methods, have been proposed for use in these two contexts. However, it seems that these second generation methodologies have received considerably less attention for smoothing splines. Thus it would be interesting to investigate the uses of these second generation methodologies in the smoothing spline setting. In below three such second generation methods will be considered. The first one is an improved version of the classical Akaike Information criterion (AIC) while the remaining two are based on a bias–variance decomposition formula of the $L_2$ risk (defined below). To the best of the author’s knowledge, the empirical performances of the last two second generation methods have never been assessed in the literature.

In Section 2 the six different smoothing parameter selection methods are reviewed. Section 3 compares these methods via a simulation study, while conclusions and recommendations are offered in Section 4.

2 Smoothing Parameter Selection Methods

First we define some notation. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{f} = (f(x_1), \ldots, f(x_n))^T$ and $\hat{\mathbf{f}}_\lambda = (\hat{f}_\lambda(x_1), \ldots, \hat{f}_\lambda(x_n))^T$. Further, let $S_\lambda$ be the "hat" matrix that maps $\mathbf{y}$ into $\hat{\mathbf{f}}_\lambda$: $\hat{\mathbf{f}}_\lambda = S_\lambda \mathbf{y}$. One can show that $S_\lambda = (I + \lambda K)^{-1}$, where $I$ is the identity matrix and $K$ is a matrix depending only on $x_1, \ldots, x_n$ (see, e.g., Green & Silverman 1994, Chapter 2). The risk
\( R(f, \hat{f}_\lambda) \) is defined as \( R(f, \hat{f}_\lambda) = \frac{1}{n} E \| f - \hat{f}_\lambda \|^2 \), where \( \| t \| = \sqrt{t^T t} \) is the usual \( L_2 \) norm for any vector \( t \). The trace of a matrix \( A \) is denoted as \( \text{tr}(A) \).

### 2.1 Classical Methods

**Cross-Validation:** Let \((S_\lambda)_{ii}\) be the \(i\)th diagonal element of \(S_\lambda\). For smoothing splines the usual leave-one-out CV score function is

\[
\text{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - (S_\lambda)_{ii}} \right\}^2,
\]

and \(\lambda\) is chosen as the minimizer of \(\text{CV}(\lambda)\).

**Generalized Cross-Validation:** The basic idea of GCV is to replace the denominators \(1 - (S_\lambda)_{ii}\) of \(\text{CV}(\lambda)\) by their average \(1 - n^{-1} \text{tr}(S_\lambda)\), giving the GCV score function

\[
\text{GCV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - n^{-1} \text{tr}(S_\lambda)} \right\}^2.
\]

As for \(\text{CV}(\lambda)\), \(\lambda\) is chosen as the minimizer of \(\text{GCV}(\lambda)\).

**Mallows' \(C_p\) Criterion:** The criterion is

\[
C_p(\lambda) = \frac{1}{n} \left\{ \| (S_\lambda - I) y \|^2 + 2 \sigma^2 \text{tr}(S_\lambda) + \sigma^2 \right\}.
\]

It is straightforward to show that \( E\{C_p(\lambda)\} = R(f, \hat{f}_\lambda) \). Unless \(\sigma^2\) is known, in practice one needs to replace the \(\sigma^2\) in \(C_p(\lambda)\) by an (independent) estimate \(\hat{\sigma}^2\) if one wants to choose \(\lambda\) as the minimizer of \(C_p(\lambda)\). In our simulation study \(\sigma^2\) is estimated by

\[
\hat{\sigma}^2_p = \frac{\| (S_{\lambda_p} - I) y \|^2}{\text{tr}(1 - S_{\lambda_p})} = \frac{\sum_{i=1}^{n} \left\{ y_i - \hat{f}_{\lambda_p}(x_i) \right\}^2}{\text{tr}(1 - S_{\lambda_p})},
\]

where \(\lambda_p\) is pre-chosen by CV. Reasons for using \(\text{tr}(1 - S_{\lambda_p})\) as the normalizing constant in \(\hat{\sigma}^2_p\) can be found for example in Green & Silverman (1994, Section 3.4).
2.2 Second Generation Methods

Improved Akaike Information Criterion: Originally for parametric problems the classical AIC was developed as an approximately unbiased estimator of the expected Kullback-Leibler information. In Hurvich, Simonoff & Tsai (1998) an improved version, AICc, of AIC is constructed for choosing the amount of smoothing for linear nonparametric smoothers. This improved criterion, for the current setting, is

\[ \text{AIC}_c(\lambda) = \log \frac{||(S_\lambda - I)y||^2}{n} + 1 + \frac{2 \{ \text{tr}(S_\lambda) + 1 \}}{n - \text{tr}(S_\lambda) - 2}. \]

It is improved in the sense that the finite sample bias of the classical AIC is corrected. As before, \( \lambda \) is chosen as the minimizer of AICc(\( \lambda \)).

Risk Estimation using Classical Pilots (RECP): A direct calculation leads to the bias–variance decomposition for \( R(\hat{f}, \hat{f}_\lambda) \):

\[ R(\hat{f}, \hat{f}_\lambda) = \frac{1}{n} E \|\hat{f} - \hat{f}_\lambda\|^2 = \frac{1}{n} \left\{ \|(S_\lambda - I)f\|^2 + \sigma^2 \text{tr}(S_\lambda S_\lambda^T) \right\}. \] (2)

The idea now is to estimate the risk \( R(\hat{f}, \hat{f}_\lambda) \) by plugging-in pilot estimates of \( f \) and \( \sigma^2 \) into (2), and choose the \( \lambda \) that minimizes the resulting risk estimator. A quick and simple strategy for choosing the pilot estimates, as suggested by Wand & Gutierrez (1997), is to use the “blocking method” of Härdle & Marron (1995). A second strategy is first to apply a classical method to select a pilot \( \lambda_p \) and then use this \( \lambda_p \) to compute \( \hat{f}_{\lambda_p} \) and \( \hat{\sigma}_p^2 \), the pilot estimates for \( f \) and \( \sigma^2 \) respectively. This second strategy of choosing the pilots has been shown to be very successful in both the local linear regression and nonparametric spectral density estimation contexts (Lee & Solo 1999 and Lee 2001). In below we shall follow this
strategy and use CV to choose $\lambda_p$. That is, the RECP method chooses the final $\lambda$ as the minimizer of $R(\hat{f}_{\lambda_p}, \hat{f}_\lambda)$, hoping that $R(\hat{f}_{\lambda_p}, \hat{f}_\lambda)$ is a good estimator for $R(f, \hat{f}_\lambda)$.

This risk estimation idea has been, sometimes under different names, applied by other authors to handle various smoothing parameter selection problems. Those different names include stabilized selectors (Chiu 1991 and Chiu 1992), smoothed cross-validation (Hall, Marron & Park 1992), double smoothing (Härdle, Hall & Marron 1992), plug-in and unbiased risk estimation (Lee & Solo 1999 and Lee 2001) and exact risk approach (Wand & Gutierrez 1997).

**Exact Double Smoothing (EDS):** Wand & Gutierrez (1997) suggested another approach, termed exact double smoothing, for choosing the pilot estimates for the risk decomposition (2). This approach involves the choosing of two “levels” of pilot estimates, and it can be briefly described as follows. Let $\lambda_0$ be the optimal $\lambda$ that minimizes $R(\hat{f}, \hat{f}_\lambda)$. One could then aim to choose the pilot smoothing parameter $\lambda_p$ as the minimizer of $E\{(\lambda_0 - \lambda)^2\}$. However, since $\lambda_0$ is an unknown, practical minimization of $E\{(\lambda_0 - \lambda)^2\}$ is not feasible. Nevertheless, Wand & Gutierrez (1997) proposed the following procedure for (approximately) carrying out such a minimization. First a closed-form approximation, $L(\lambda_0, \lambda)$, for $E\{(\lambda_0 - \lambda)^2\}$ is derived. Then these authors suggest replacing the unknown $\lambda_0$ with a second pilot smoothing parameter $\lambda_p$. That is, the unknown $E\{(\lambda_0 - \lambda)^2\}$ is now approximated by the computable quantity $L(\lambda_p, \lambda)$. The derived expression for $L(\lambda_p, \lambda)$ is

$$L(\lambda_p, \lambda) = \{f^{T}_{\lambda_p} D_{\lambda_p} f_{\lambda_p} + \delta_{\lambda_p}^2 \text{tr}(D_{\lambda_p})\}^2 + 4\delta_{\lambda_p}^2 f^{T}_{\lambda_p} D_{\lambda_p}^2 f_{\lambda_p} + 2\delta_{\lambda_p}^4 \text{tr}(D_{\lambda_p}^2)$$

$$+ 4\delta_{\lambda_p}^2 \text{tr}(S_{\lambda_p}^T S_{\lambda_p}) \times \{f^{T}_{\lambda_p} D_{\lambda_p} f_{\lambda_p} + \delta_{\lambda_p}^2 \text{tr}(D_{\lambda_p})\} + 4\delta_{\lambda_p}^4 \text{tr}(S_{\lambda_p}^T S_{\lambda_p}^T),$$
where $S'_{\lambda_{p_2}} = \lambda_{p_2}^{-1} S_{\lambda_{p_2}} (S_{\lambda_{p_2}} - I)$ and $D_{\lambda\lambda_{p_2}} = 2\lambda_{p_2}^{-1} S_{\lambda} S_{\lambda_{p_2}} (S_{\lambda_{p_2}} - I)^2 S_{\lambda}$.

To sum up, this EDS method chooses $\lambda$ as the minimizer of $R(\hat{f}_{\lambda_{p_1}}, \hat{f}_{\lambda})$, where $\lambda_{p_1}$ minimizes $L(\lambda_{p_2}, \lambda)$. In our simulation we choose $\lambda_{p_2}$ using CV.

We recall again that for RECP and EDS, no assessment of their practical performances have been reported in the literature (not even in Wand & Gutierrez 1997).

### 2.3 Speed Comparisons

The three methods CV, GCV and AIC$_C$ require roughly the same amount of computational time for obtaining their corresponding $\lambda$, as their computations only involve one (nearly identical) numerical minimization. Compare to these three methods, both $C_p$ and RECP require a longer computational time, as there are two numerical minimizations involved. However notice that some calculations are redundant for these two numerical minimizations, and therefore the overall computational time will not be doubled if one exercises careful programming. Lastly the computation of EDS involves three minimizations, but again some calculations are redundant and hence the overall time will not be tripled.

### 3 Simulation Study

This section reports the results of a simulation study that was conducted to evaluate the performances of the above six smoothing parameter selection methods. The adopted experimental setup was essentially the same as in Wand (2000). This setup, originally due to Professor Steve Marron, was designed to study the effects of varying the (i) noise level, (ii)
design density, (iii) degree of spatial variation and (iv) noise variance function in an independent and effective fashion. The idea is as follows. Totally four sets of numerical experiments are to be performed. For each set of experiments, only one of the above four experimental factors (e.g., noise level) is changed while the remaining three are being kept unchanged. Within each set of experiments, the factor under consideration is changed six times, and hence there are altogether 24 different configurations. In this way it is believed that patterns can be more easily detected. The only change that we have made to this setup of Wand (2000) was that, for the spatial variation factor, $n = 200$ was used instead of $n = 400$. The number of replications for each of the 24 configurations were 200. For completeness, the setup specification is listed in Table 1. The software used was S-Plus and the algorithm described in Green & Silverman (1994, Chapter 2) was used to compute $\hat{f}_\lambda$ for a given $\lambda$.

For each simulated data set, the numerical measure that was used to evaluate the quality of any selected $\lambda'$ is

$$r = \frac{R(f, \hat{f}_\lambda)}{\min_\lambda R(f, \hat{f}_\lambda)} \geq 1.$$  

That is, the smaller the $r$ value, the better the quality of $\lambda'$. Boxplots of the log$_e$ $r$ values for the 24 different configurations are given in Figures 1 to 4.

Paired Wilcoxon tests were also applied to test if the difference between the median $r$ values of any two methods is significant or not. The significance level used was 5% = 0.83%. The methods were also ranked in the following manner. If the median $r$ value of a method is significantly less than the remaining five, it will be assigned a rank 1. If the median $r$ value of a method is significantly larger than one but less than four methods, it will be assigned
a rank 2, and similarly for ranks 3 to 6. Methods having non-significantly different median values will share the same averaged rank. The resulting rankings are also given in Figures 1 to 4, and the averaged rankings are tabulated in Table 2.

4 Conclusions and Recommendations

The overall Wilcoxon test rankings for CV, GCV, $C_p$, AICc, RECP and EDS are, respectively, 3.86, 3.67, 3.67, 3.40, 2.96 and 3.46. Therefore judging by this measure it may seem that RECP is the best method. However, this statement should not be blindly believed, as firstly it is only based on results from simulations, and, secondly, RECP was not uniformly better than the other methods in the 24 different simulation configurations.

From a closer inspection of the simulation results, the following observations were made.

- No method performed uniformly the best.

- In most cases the six methods actually gave very reasonable performances, say with a log$_e r$ value less than 0.5.

- The three classical methods, CV, GCV and $C_p$, gave very similar results. In fact for all 24 simulation configurations GCV and $C_p$ always shared the same ranking.

- The second generation method, AICc, has never given a worse performance than CV, GCV or $C_p$.

- For a simple regression function with a high noise level, the two second generation risk-based methods (RECP and EDS) seem to be superior; see Figure 1, $j = 5$ and $j = 6$. 

• All methods shared the same ranking for those design density factor experiments.

• Under heteroskedastic errors AICc seems to be a better method; see Figure 4, \( j = 1 \) and \( j = 3 \).

*Which method to use?* To answer this question we first recall that no method performed uniformly the best. Therefore one may use other criteria, such as speed, to help selecting a method. Our recommendation is as follows. If speed is not an issue and if, say checked by visual inspection, the homoskedastic error assumption is satisfied, use RECP; otherwise use the faster AICc.
References


Figure 1: Results for changing the noise level factor. In each pair of panels the left half displays one typical simulated data set together with the true regression function. The right half are the boxplots of the log₂ r values for, from left to right, CV, GCV, C_p, AICc, RECP and EDS. The numbers below the boxplots are the paired Wilcoxon test rankings.
Figure 2: Similar to Figure 1 but for the design density factor.
Figure 3: Similar to Figure 1 but for the spatial variation factor.
Figure 4: Similar to Figure 1 but for the variance function factor.
\begin{tabular}{|c|c|c|c|c|}
\hline
factor & generic form & particular choices \\
\hline
noise level & $y_{ij} = f(x_i) + \sigma_j \epsilon_i$ & $\sigma_j = 0.02 + 0.04(j - 1)^2$ \\
\hline
design density & $y_{ij} = f(x_{ji}) + \sigma \epsilon_i$ & $\sigma = 0.1$, $X_{ji} = F_j^{-1}(x_i)$ \\
\hline
spatial variation & $y_{ij} = f_j(x_i) + \sigma \epsilon_i$ & $\sigma = 0.2$, $f_j(x) = \sqrt{x(1-x)} \sin \left( \frac{2\pi(1+2^\left(\frac{j-1}{2}\right))}{x+2^\left(\frac{j-1}{2}\right)} \right)$ \\
\hline
variance function & $y_{ij} = f(x_i) + \sqrt{v_j(x_i)} \epsilon_i$ & $v_j(x) = [0.15(1+0.4(2j-7)(x-0.5))]^2$ \\
\hline
\end{tabular}

\begin{itemize}
\item $j = 1, \ldots, 6$; \quad $n = 200$; \quad $x_i = \frac{i-0.5}{n}$; \quad $\epsilon_i \sim$ iid $N(0,1)$
\item $f(x) = 1.5 \phi \left( \frac{x-0.35}{0.15} \right) - \phi \left( \frac{x-0.8}{0.04} \right)$; \quad $\phi(u) = \frac{1}{\sqrt{2\pi}} \exp \left( \frac{-u^2}{2} \right)$
\item $X_i \sim$ iid Uniform$[0,1]$; \quad $F_j$ is the Beta $\left( \frac{i+4}{5}, \frac{11-i}{5} \right)$ c.d.f.
\end{itemize}

Table 1: Specification of the simulation setup.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
 & Noise Level & Design Density & Spatial Variation & Variance Function & Overall \\
\hline
CV & 4.17 & 3.50 & 3.75 & 4.00 & 3.86 \\
\hline
GCV & 3.67 & 3.50 & 3.75 & 3.75 & 3.67 \\
\hline
$C_p$ & 3.67 & 3.50 & 3.75 & 3.75 & 3.67 \\
\hline
AICc & 3.67 & 3.50 & 3.75 & 3.33 & 3.40 \\
\hline
RECP & 2.25 & 3.50 & 2.75 & 3.50 & 2.96 \\
\hline
EDS & 3.58 & 3.50 & 3.25 & 3.50 & 3.46 \\
\hline
\end{tabular}
\caption{Averaged Wilcoxon test rankings for the six smoothing parameter selection methods.}
\end{table}