ELSEVIER Computational Statistics & Data Analysis 42 (2003) 139–148

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Smoothing parameter selection for smoothing splines: a simulation study

Thomas C.M. Lee*

Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA

Received 1 June 2001; received in revised form 1 May 2002

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 risk estimation methods. To the best of the author's knowledge, the empirical performances of these two risk estimation methods have never been reported in the literature. Empirical conclusions from and recommendations based on the simulation results will be provided. One noteworthy empirical observation is that the popular method, generalized cross-validation, was outperformed by another method, an improved Akaike Information criterion, that shares the same assumptions and computational complexity.

Keywords: Exact double smoothing; Nonparametric regression; Plug-in methods; Risk estimation; Roughness penalty; Smoothing parameter; Smoothing splines

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

$$y_i = f(x_i) + \varepsilon_i, \quad a < x_1 < \cdots < x_n < b, \quad \varepsilon_i \sim \text{ iid } \mathrm{N}(0, \sigma^2),$$

* Tel.: +970-491-2185; fax: +970-491-7895.
E-mail address: tlee@stat.colostate.edu (T.C.M. Lee).

0167-9473/02/$ - see front matter © 2002 Published by Elsevier Science B.V.
PII: S0167-9473(02)00159-7

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

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

In the above λ 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 example, Eubank (1988); Green and Silverman (1994) and Wahba (1990).

It is widely known that λ has a crucial effect on the quality of $\hat{f}_\lambda$. Therefore, obtaining an appropriate choice of λ 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 λ. 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. Four of them are "classical" while the remaining two are so-called *risk estimation* methods. The four classical methods are cross-validation (CV), generalized cross-validation (GCV), Mallows' C_p criterion and an improved version of the classical Akaike Information criterion (AIC). In both the local linear regression and kernel density estimation contexts, it has been known that many classical selectors tend to be highly variable and also have the tendency to undersmooth. In order to repair these drawbacks, various new methods, such as risk estimation or "plug-in" methods, have been proposed for the use in these two contexts. However, it seems that these new methodologies have received considerably less attention for smoothing splines. Thus, it would be interesting to investigate the use of these new methodologies in the smoothing spline setting. Below two new risk estimation methods will be considered. They both 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 these two risk estimation 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 $\boldsymbol{y}=(y_1,\ldots,y_n)^{\mathrm{T}}$, $\boldsymbol{f}=(f(x_1),\ldots,f(x_n))^{\mathrm{T}}$ and $\hat{\boldsymbol{f}}_\lambda=(\hat{f}_\lambda(x_1),\ldots,\hat{f}_\lambda(x_n))^{\mathrm{T}}$. Further, let S_λ be the "hat" matrix that maps $\boldsymbol{y}$ into $\hat{\boldsymbol{f}}_\lambda$: $\hat{\boldsymbol{f}}_\lambda = S_\lambda \boldsymbol{y}$. 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 and Silverman, 1994, Chapter 2). The risk $R(\boldsymbol{f},\hat{\boldsymbol{f}}_\lambda)$ is defined as $R(\boldsymbol{f},\hat{\boldsymbol{f}}_\lambda)=1/nE\|\boldsymbol{f}-\hat{\boldsymbol{f}}_\lambda\|^2$, where $\|\boldsymbol{t}\|=\sqrt{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{t}}$ is the usual L_2 norm for any vector $\boldsymbol{t}$. The trace of a matrix A is denoted as $\mathrm{tr}(A)$.

2.1. Classical methods

Cross-validation. Let $(S_\lambda)_{ii}$ be the ith diagonal element of S_λ. For smoothing splines the usual leave-one-out CV score function is

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

and λ is chosen as the minimizer of $\mathrm{CV}(\lambda)$.

Generalized cross-validation. The basic idea of GCV is to replace the denominators $1-(S_\lambda)_{ii}$ of $\mathrm{CV}(\lambda)$ by their average $1-n^{-1}\operatorname{tr}(S_\lambda)$, giving the GCV score function

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n}\{y_i - \hat{f}_\lambda(x_i)\}^2}{\{1-n^{-1}\operatorname{tr}(S_\lambda)\}^2}.$$

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

Mallows' C_p criterion. The criterion is

$$C_p(\lambda) = \frac{1}{n}\{\|(S_\lambda - I)\boldsymbol{y}\|^2 + 2\sigma^2\operatorname{tr}(S_\lambda) + \sigma^2\}.$$

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

$$\hat{\sigma}^2_{\lambda_p} = \frac{\|(S_{\lambda_p} - I)\boldsymbol{y}\|^2}{\operatorname{tr}(1-S_{\lambda_p}} = \frac{\sum_{i=1}^{n}\{y_i - \hat{f}_{\lambda_p}(x_i)\}^2}{\operatorname{tr}(1-S_{\lambda_p})}, \tag{1}$$

where λ_p is pre-chosen by CV. Reasons for using $\operatorname{tr}(1-S_{\lambda_p})$ as the normalizing constant in $\hat{\sigma}^2_{\lambda_p}$ can be found for example in Green and Silverman (1994, Section 3.4).

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 et al. (1998) an improved version, $\mathrm{AIC_c}$, of AIC is constructed for choosing the amount of smoothing for linear nonparametric smoothers. This improved criterion, for the current setting, is

$$\mathrm{AIC_c}(\lambda) = \log\frac{\|(S_\lambda - I)\boldsymbol{y}\|^2}{n} + 1 + \frac{2\{\operatorname{tr}(S_\lambda)+1\}}{n-\operatorname{tr}(S_\lambda)-2}.$$

It is improved in the sense that the finite sample bias of the classical AIC is corrected. As before, λ is chosen as the minimizer of $\mathrm{AIC_c}(\lambda)$.

2.2. Risk estimation methods

This section describes two risk estimation methods for choosing λ. These two methods, similar to some traditional "plug-in" methods for smoothing parameter selection (e.g., Ruppert et al., 1995 for local linear smoothing), require the choosing of some *pilot estimates* (see below). However, unlike those traditional "plug-in" methods, these two risk estimation methods do not require the existence of f'' nor an analytic

expression for the optimal smoothing parameter. For a discussion on those traditional "plug-in" methods, see, for example, Loader (1999).

2.2.1. Risk estimation using classical pilots (RECP)

A direct calculation leads to the bias-variance decomposition for $R(\boldsymbol{f},\hat{\boldsymbol{f}}_\lambda)$:

$$R(\boldsymbol{f},\hat{\boldsymbol{f}}_\lambda)=\frac{1}{n}E\|\boldsymbol{f}-\hat{\boldsymbol{f}}_\lambda\|^2=\frac{1}{n}\{\|(S_\lambda-I)\boldsymbol{f}\|^2+\sigma^2\,\mathrm{tr}(S_\lambda S_\lambda^{\mathrm{T}})\}. \tag{2}$$

The idea now is to estimate the risk $R(\boldsymbol{f},\hat{\boldsymbol{f}}_\lambda)$ by plugging-in pilot estimates of $\boldsymbol{f}$ and σ^2 into (2), and choose the λ that minimizes the resulting risk estimator. A quick and simple strategy for choosing the pilot estimates, as suggested by Wand and Gutierrez (1997), is to use the "blocking method" of Härdle and Marron (1995). A second strategy is first to apply a classical method to select a pilot λ_p and then use this λ_p to compute $\hat{\boldsymbol{f}}_{\lambda_p}$ and $\hat{\sigma}^2_{\lambda_p}$, the pilot estimates for $\boldsymbol{f}$ and σ^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 and Solo, 1999; Lee, 2001). In below we shall follow this strategy and use CV to choose λ_p. That is, the RECP method chooses the final λ as the minimizer of $R(\hat{\boldsymbol{f}}_{\lambda_p},\hat{\boldsymbol{f}}_\lambda)$, hoping that $R(\hat{\boldsymbol{f}}_{\lambda_p},\hat{\boldsymbol{f}}_\lambda)$ is a good estimator for $R(\boldsymbol{f},\hat{\boldsymbol{f}}_\lambda)$.

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; Chiu, 1992), smoothed cross-validation (Hall et al., 1992), double smoothing (Härdle et al., 1992), plug-in and unbiased risk estimation (Lee and Solo, 1999; Lee, 2001) and exact risk approach (Wand and Gutierrez, 1997).

2.2.2. Exact double smoothing (EDS)

Wand and 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 λ_0 be the optimal λ that minimizes $R(\boldsymbol{f},\hat{\boldsymbol{f}}_\lambda)$. One could then, given the data, aim to choose the pilot smoothing parameter λ_{p_1} as the minimizer of $E\{(\lambda_0-\lambda)^2\}$. However, since λ_0 is an unknown, practical minimization of $E\{(\lambda_0-\lambda)^2\}$ is not feasible. Nevertheless, Wand and 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 λ_0 with a second pilot smoothing parameter λ_{p_2}. That is, the unknown $E\{(\lambda_0-\lambda)^2\}$ is now approximated by the computable quantity $L(\lambda_{p_2},\lambda)$. The derived expression for $L(\lambda_{p_2},\lambda)$ is

$$\begin{aligned}L(\lambda_{p_2},\lambda)=&\{\hat{\boldsymbol{f}}^{\mathrm{T}}_{\lambda_{p_2}}D_{\lambda\lambda_{p_2}}\hat{\boldsymbol{f}}_{\lambda_{p_2}}+\hat{\sigma}^2_{\lambda_{p_2}}\,\mathrm{tr}(D_{\lambda\lambda_{p_2}})\}^2+4\hat{\sigma}^2_{\lambda_{p_2}}\hat{\boldsymbol{f}}^{\mathrm{T}}_{\lambda_{p_2}}D^2_{\lambda\lambda_{p_2}}\hat{\boldsymbol{f}}_{\lambda_{p_2}}\\&+2\hat{\sigma}^4_{\lambda_{p_2}}\,\mathrm{tr}(D^2_{\lambda\lambda_{p_2}})+4\hat{\sigma}^2_{\lambda_{p_2}}\,\mathrm{tr}(S'_{\lambda_{p_2}}S^{\mathrm{T}}_{\lambda_{p_2}})\{\hat{\boldsymbol{f}}^{\mathrm{T}}_{\lambda_{p_2}}D_{\lambda\lambda_{p_2}}\hat{\boldsymbol{f}}_{\lambda_{p_2}}\\&+\hat{\sigma}^2_{\lambda_{p_2}}\,\mathrm{tr}(D_{\lambda\lambda_{p_2}})\}+4\hat{\sigma}^4_{\lambda_{p_2}}\,\mathrm{tr}^2(S'_{\lambda_{p_2}}S^{\mathrm{T}}_{\lambda_{p_2}}),\end{aligned}$$

where $S'_{\lambda_{p_2}}=\lambda_{p_2}^{-1}S_{\lambda_{p_2}}(S_{\lambda_{p_2}}-I)$ and $D_{\lambda\lambda_{p_2}}=2\lambda_{p_2}^{-1}S_\lambda S_{\lambda_{p_2}}(S_{\lambda_{p_2}}-I)^2S_\lambda$.

To sum up, this EDS method chooses λ as the minimizer of $R(\hat{\boldsymbol{f}}_{\lambda_{p_1}}, \hat{\boldsymbol{f}}_{\lambda})$, where λ_{p_1} minimizes $L(\lambda_{p_2}, \lambda)$. In our simulation we choose λ_{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 and Gutierrez, 1997).

2.3. Speed comparisons

The three methods CV, GCV and $\mathrm{AIC_c}$ require roughly the same amount of computational time for obtaining their corresponding λ, 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 and Silverman (1994, Chapter 2) was used to compute $\hat{\boldsymbol{f}}_{\lambda}$ for a given λ.

For each simulated data set, the numerical measure that was used to evaluate the quality of any selected λ' is the ratio r of two mean-squared errors:

$$r = \frac{\|\boldsymbol{f} - \hat{\boldsymbol{f}}_{\lambda'}\|^2}{\min_{\lambda} \|\boldsymbol{f} - \hat{\boldsymbol{f}}_{\lambda}\|^2} \geqslant 1.$$

That is, the smaller the r value, the better the quality of λ'. Boxplots of the $\log_e r$ values for the 24 different configurations are given in Figs. 1–4.

Table 1
Specification of the simulation setup

Factor	Generic form	Particular choices
Noise level	$y_{ij} = f(x_i) + \sigma_j \varepsilon_i$	$\sigma_j = 0.02 + 0.04(j-1)^2$
Design density	$y_{ij} = f(X_{ji}) + \sigma \varepsilon_i$	$\sigma = 0.1,\ X_{ji} = F_j^{-1}(X_i)$
Spatial variation	$y_{ij} = f_j(x_i) + \sigma \varepsilon_i$	$\sigma = 0.2, f_j(x) = \sqrt{x(1-x)}\sin[\frac{2\pi\{1+2^{(9-4j)/5}\}}{x+2^{(9-4j)/5}}]$
Variance function	$y_{ij} = f(x_i) + \sqrt{v_j(x_i)}\varepsilon_i$	$v_j(x) = [0.15\{1 + 0.4(2j-7)(x-0.5)\}]^2$

$j = 1, \ldots, 6; \quad n = 200; \quad x_i = \frac{i-0.5}{n}; \quad \varepsilon_i \sim$ iid $N(0,1)$

$f(x) = 1.5\phi(\frac{x-0.35}{0.15}) - \phi(\frac{x-0.8}{0.04}); \quad \phi(u) = \frac{1}{\sqrt{2\pi}}\exp(\frac{-u^2}{2})$

$X_i \sim$ iid Uniform$[0,1]$; $F_j$is the Beta$(\frac{j+4}{5}, \frac{11-j}{5})$ c.d.f.

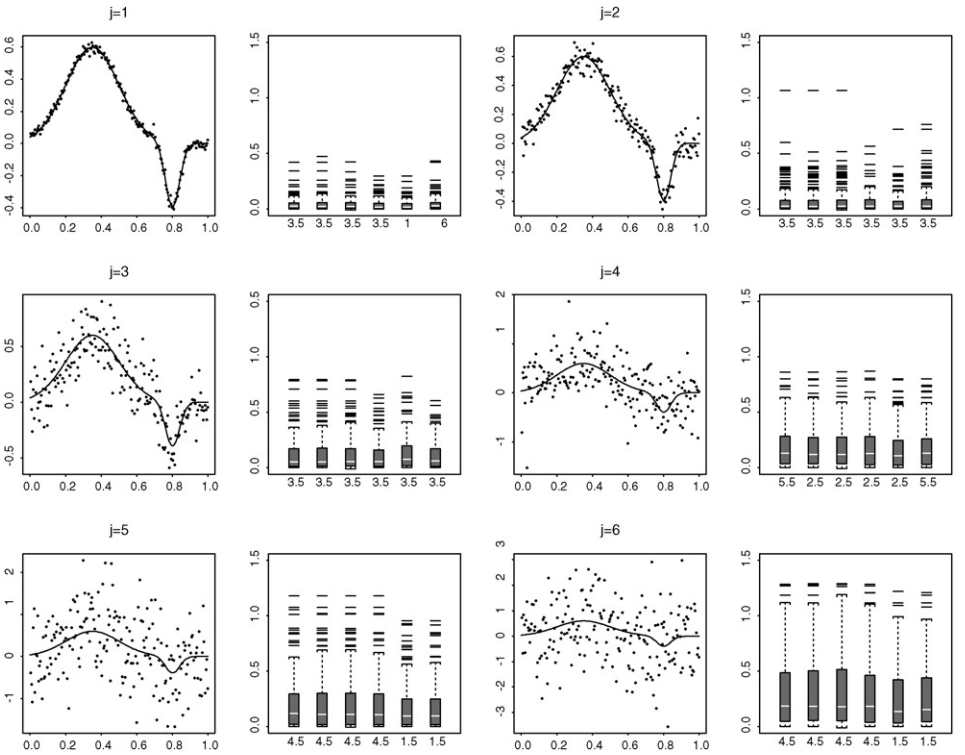

Fig. 1. Results for changing the noise level factor. In each pair of panels the left-half displays one typical simulated data set together with the true regression function. The right-half are the boxplots of the $\log_e r$ values for, from left to right, CV, GCV, C_p, AIC_c, RECP and EDS. The numbers below the boxplots are the paired Wilcoxon test rankings.

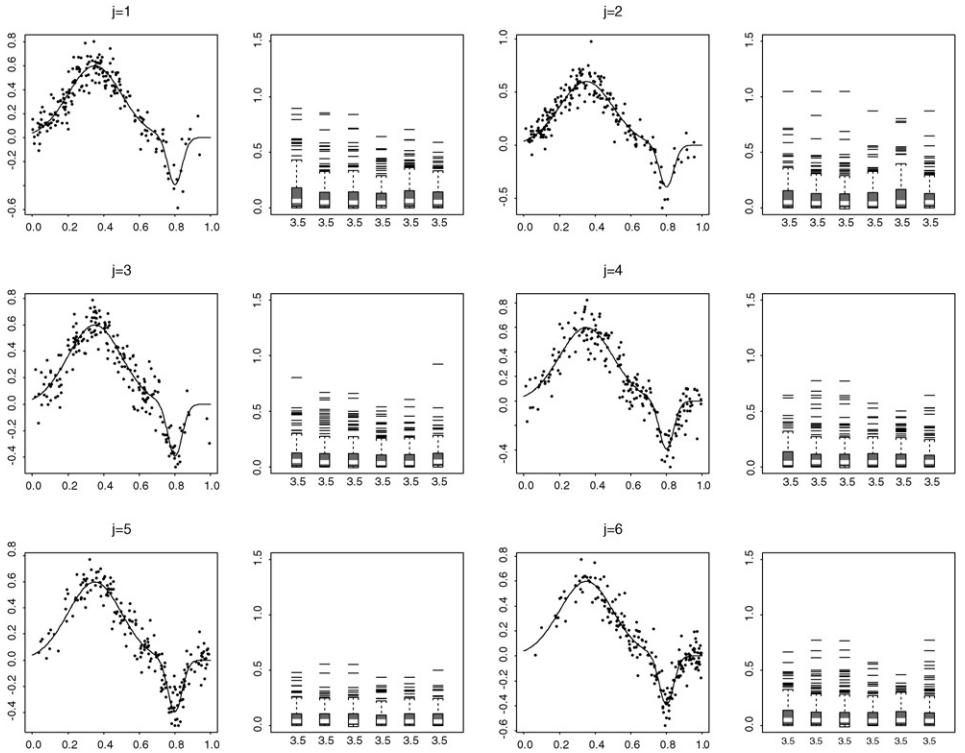

Fig. 2. Similar to Fig. 1 but for the design density factor.

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 $\frac{5}{6}\% = 0.83\%$. 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 3–6. Methods having non-significantly different median values will share the same averaged rank. The resulting rankings are also given in Figs. 1–4, and the averaged rankings are tabulated in Table 2.

4. Conclusions and recommendations

The overall Wilcoxon test rankings for CV, GCV, C_p, $\mathrm{AIC_c}$, 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,

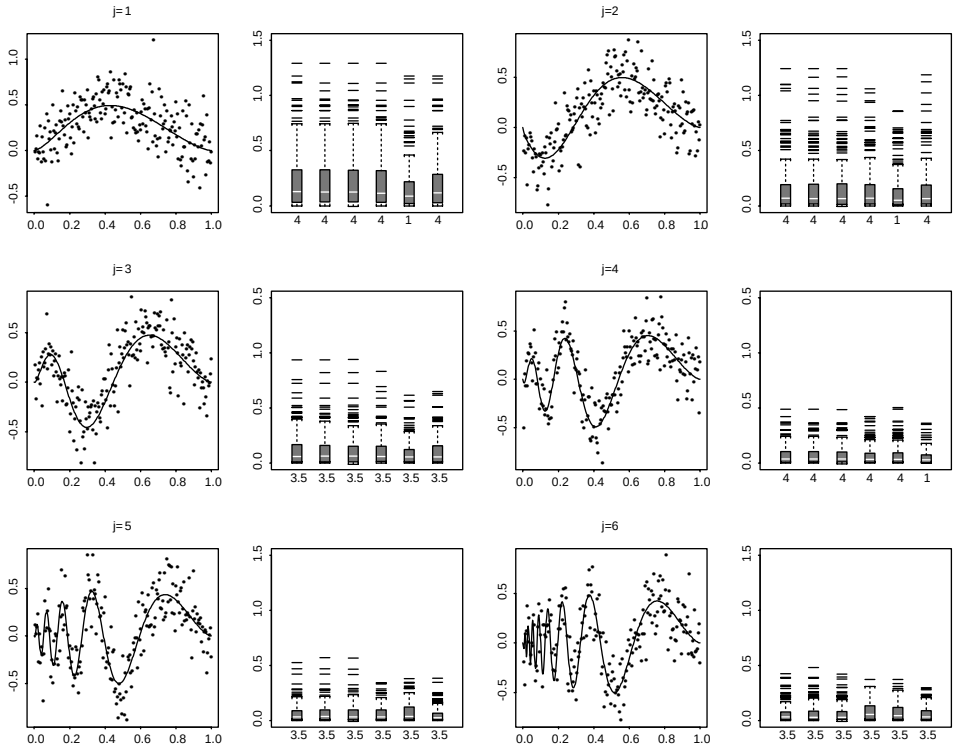

Fig. 3. Similar to Fig. 1 but for the spatial variation factor.

Table 2
Averaged Wilcoxon test rankings for the six smoothing parameter selection methods

	Noise level	Design density	Spatial variation	Variance function	Overall
CV	4.17	3.50	3.75	4.00	3.86
GCV	3.67	3.50	3.75	3.75	3.67
C_p	3.67	3.50	3.75	3.75	3.67
AIC_c	3.67	3.50	3.75	2.67	3.40
RECP	2.25	3.50	2.75	3.33	2.96
EDS	3.58	3.50	3.25	3.50	3.46

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.

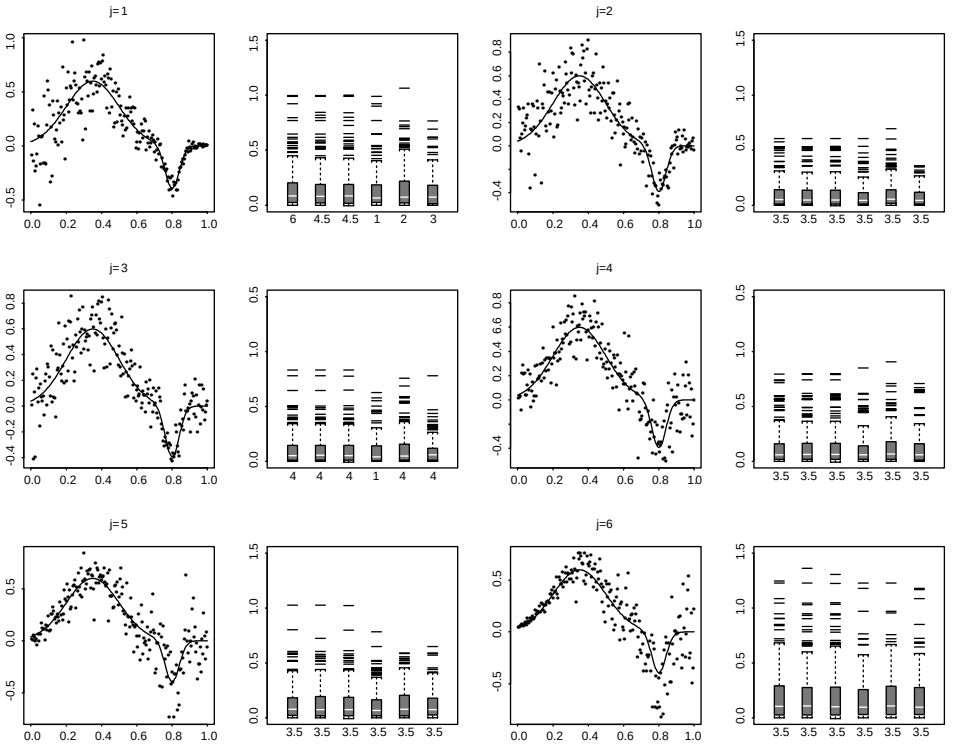

Fig. 4. Similar to Fig. 1 but for the variance function factor.

- 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 classical method, AIC_c, has never given a worse performance than CV, GCV or C_p.
- For a simple regression function with a high noise level, the two risk estimation methods (RECP and EDS) seem to be superior; see Fig. 1, $j = 5$ and 6.
- All methods shared the same ranking for those design density factor experiments.
- Under heteroskedastic errors AIC_c seems to be a better method; see Fig. 4, $j = 1$ and 3. It is probably because AIC_c is targeting at the expected Kullback–Leibler discrepancy which considers the distance for the whole error distribution, while the remaining methods are targeting at the L_2 risk which only considers the mean.

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 AIC_c.

Acknowledgements

The author would like to thank the referees for their very constructive comments.

References

Chiu, S.-T., 1991. Some stabilized bandwidth selectors for nonparametric regression. Ann. Statist. 19, 1528–1546.

Chiu, S.-T., 1992. An automatic bandwidth selector for kernel density estimation. Biometrika 79, 771–782.

Eubank, R.L., 1988. Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York.

Green, P.J., Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models. Chapman & Hall, London.

Hall, P., Marron, J.S., Park, B.U., 1992. Smoothed cross-validation. Probab. Theory Related Fields 92, 1–20.

Härdle, W., Marron, J.S., 1995. Fast and simple scatterplot smoothing. Comput. Statist. Data Anal. 20, 1–17.

Härdle, W., Hall, P., Marron, J.S., 1992. Regression smoothing parameters that are not far from their optimum. J. Amer. Statist. Assoc. 87, 227–233.

Hurvich, C.M., Simonoff, J.S., Tsai, C.-L., 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. Roy. Statist. Soc. Ser. B 60, 271–293.

Lee, T.C.M., 2001. A stabilized bandwidth selection method for kernel smoothing of the periodogram. Signal Process. 81, 419–430.

Lee, T.C.M., Solo, V., 1999. Bandwidth selection for local linear regression: a simulation study. Comput. Statist. 14, 515–532.

Loader, C., 1999. Local Regression and Likelihood. Springer, New York.

Ruppert, D., Sheather, S.J., Wand, M.P., 1995. An effective bandwidth selector for local least squares regression. J. Amer. Statist. Assoc. 90, 1257–1270.

Wahba, G., 1990. Spline models for observational data. CBMS-NSF, Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.

Wand, M.P., 2000. A comparison of regression spline smoothing procedures. Comput. Statist. 15, 443–462.

Wand, M.P., Gutierrez, R.G., 1997. Exact risk approaches to smoothing parameter selection. J. Nonparametric Statist. 8, 337–354.