Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: http://www.tandfonline.com/loi/gnst20

Exact risk approaches to smoothing parameter selection

M. P. Wand & R. G. Gutierrez

To cite this article: M. P. Wand & R. G. Gutierrez (1997) Exact risk approaches to smoothing parameter selection, Journal of Nonparametric Statistics, 8:4, 337-354, DOI: 10.1080/10485259708832729

To link to this article: http://dx.doi.org/10.1080/10485259708832729

Nonparametric Statistics, Vol. 8, pp. 337–354
Reprints available directly from the publisher
Photocopying permitted by license only

EXACT RISK APPROACHES TO SMOOTHING PARAMETER SELECTION

M. P. WAND[a] and R. G. GUTIERREZ[b]

[a]*Australian Graduate School of Management, University of New South Wales, Sydney, 2052, Australia;* [b]*Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332, USA*

(Received 14 May 1997; In final form 3 June 1997)

The past decade has seen the development of a large number of second-generational smoothing parameter selectors as a response to the high degree of variability of cross-validatory methods. However, most of these rules rely on asymptotic approximations which make them subject to adverse performance when the approximations are poor. They are also difficult to extend to those settings where asymptotics is difficult. We aim to alleviate each of these problems by developing rules based on exact expressions for the risk.

Keywords: Bandwidth selection; correlated errors; cross-validation; double smoothing; kernel estimator; partially linear model; smoothing spline

1. INTRODUCTION

An analyst using a smoothing technique and whom would like to have the smoothing parameter chosen by the data is usually faced with a choice between two types of methodology. The first type comprises selection procedures that are founded upon classical prediction and model selection ideas, such as cross-validation and Mallows' unbiased risk criterion (Mallows, 1973). Several others are listed in Härdle, Hall and Marron (1988). They have the appeal of being fully automatic, except for the possible estimation of nuisance parameters such as the residual variance. They are also asymptotically equivalent to one another (Härdle, Hall and Marron, 1988) so we will refer to them simply as cross-validatory smoothing parameter selectors.

The main alternative to cross-validation is to use a second generational smoothing parameter selector. By 'second generational' we mean those selectors that have been developed over the past decade or so, fuelled by the realisation that cross-validation is subject to a high degree of sample variability. Quantifications of this shortcoming of cross-validation and a survey of second generational smoothing parameter selectors that aim for improved performance is given by Jones, Marron and Sheather (1996).

The reduction in sample variability that is usually enjoyed by second generation rules comes at price. The two main practical costs that have to be borne are:

(a) an initial 'pilot' estimate of the function being estimated is required.
(b) a relatively detailed asymptotic analysis to specify the rule or to work out the optimal choices of auxiliary smoothing parameters. The dependence on asymptotics also renders such rules to poor performance when the asymptotic approximation is poor.

Cost (a) is difficult, perhaps impossible, to escape from in theory-driven approaches to smoothing parameter selection. It appears that one cannot improve upon cross-validation without having some type of pilot estimation. The second cost has led to a lot of interesting, albeit somewhat complicated, procedures over the past few years. Examples include the rules of Jones, Marron and Park (1991), Sheather and Jones (1991), Hall, Marron and Park (1992), Härdle, Hall and Marron (1992) and Kim, Park and Marron (1994). However, almost all of them have been restricted to simple smoothers such as kernel density estimators. The main reason for this is that the required mathematics becomes either very complicated, or even intractable, for more complicated smoothing techniques such as smoothing splines and non-linear kernel smoothers. Such rules can also suffer from their dependence on asymptotic approximations which can be quite poor, particularly near boundaries.

In this paper we present a general principle called *exact double smoothing* which aims to alleviate Cost (b). The idea is to work with exact expressions as much as possible. Traditional 'small smoothing parameter' asymptotics are not used at all, which means that the rules are easier to specify and are more robust against poor asymptotic

approximations. The approach also allows the specification of second generation rules in those contexts where lack of asymptotic theory has hindered their development, such as spline smoothing and partially linear models. A final advantage of exact double smoothing is that dependence on the unknown function is explicit, rather than through higher order derivatives as is typical for current second generational selectors. This allows simpler use of pilot estimators.

Section 2 gives a general description of the proposed exact risk approaches to smoothing parameter selection. Illustrations are given in Section 3. We discuss the choice initial estimate in Section 4. Simulation results are given in Section 5 and concluding remarks are made in Section 6.

2. EXACT RISK SMOOTHING PARAMETER SELECTION

Suppose that θ is the object of interest. Typically θ would consist of values of a function, such as a regression mean function or a probability density, but it might also be an ordinary parameter in a semiparametric model. Examples of each type are given in Section 3. Let

$$\{\hat{\theta}_h : h \in \mathcal{H}\}$$

be a class of nonparametric estimates of θ where h is a smoothing parameter and $\mathcal{H}$ is the set of possible values of h. The quality of $\hat{\theta}$ as an estimate of θ will be measured by a risk $R = R(h, \theta)$. With respect to R the optimal smoothing parameter is

$$h_\theta = \operatorname*{argmin}_h R(h, \theta)$$

where minimisation is taken to be over $h \in \mathcal{H}$.

At this point it is helpful to introduce a concrete example. Consider the simple nonparametric regression model

$$Y = m + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I) \tag{1}$$

where $Y = (Y_1, \ldots, Y_n)^T$, $m = (m_1, \ldots, m_n)^T$, and $m_i = m(x_i)$ for some unknown function m and locations x_i. For simplicity we will assume

that σ^2 is known. A important class of estimates of m is those of the form

$$\hat{m}_h = S_h Y$$

where S_h is an $n \times n$ matrix, usually called the smoother matrix, parameterised by the smoothing parameter h. Examples of linear smoothers include regression splines, smoothing splines and kernel-type estimators. A common measure of the risk of $\hat{m}_h$ as an estimate of m is the mean summed squared error (MSSE)

$$R(h, m) = E\|\hat{m}_h - m\|^2$$

where $\|x\|^2 = x^T x$ for a d-variate vector x. An attractive feature of $R(h, m)$ is that it admits the closed form representation

$$R(h, m) = \|(S_h - I)m\|^2 + \sigma^2 \mathrm{tr}(S_h S_h^T).$$

Most classical cross-validatory selectors target the minimiser of this particular risk. The Mallows' criterion is

$$C_L(h) = \|(S_h - I)Y\|^2 + 2\sigma^2 \mathrm{tr}(S_h) + \sigma^2.$$

and is an unbiased estimate of $R(h, m)$. The ordinary cross-validation criterion function can be defined by

$$\mathrm{CV}(h) = \sum_{i=1}^{n} \left[\frac{\{(S_h - I)Y\}_i}{1 - (S_h)_{ii}} \right]^2.$$

For common linear smoothers such as kernel estimators and smoothing splines this is equivalent to a 'leave-one-out' sum of squares, $\sum_{i=1}^{n} \{Y_i - \hat{m}^{-i}(x_i)\}^2$, where $\hat{m}^{-i}$ is based on the data with (x_i, Y_i) excluded.

2.1. Level Zero Exact Risk Selectors

Suppose that θ_{init} is an initial estimate of θ. Then we define the corresponding level zero selector of h_θ to be

$$\hat{h}_0 = \underset{h}{\mathrm{argmin}}\, R(h, \theta_{\mathrm{init}}).$$

The initial estimate is such that it does not require a smoothing parameter to be chosen. However, it may depend on some auxiliary parameters. The simplest example of a level zero exact risk selector is a slight variant of Scott's (1979) rule for choosing the binwidth of a histogram. Let f denote the true density and $\hat{f}_h$ be the relative frequency histogram with bin edges $\{hj : j \text{ an integer}\}$. Then, with respect to the exact mean integrated squared error risk, Scott's rule is

$$\hat{h}_0 = \underset{h}{\operatorname{argmin}}\, E\int (\hat{f}_h - f_{\text{init}})^2$$

where f_{init} is a normal density with variance replaced by the sample variance s^2. The only difference between $\hat{h}_0$ and Scott's actual proposal is that he uses an asymptotic approximation to the risk, which results in the closed form binwidth rule $3.49sn^{-1/3}$. In Scott (1992) this rule is adapted by extending the class of initial estimates to handle skewness and kurtosis.

Other examples of level zero selection procedures based on asymptotic approximations to L_2-type risks include Scott's (1985) rule for the binwidth of a frequency polygon, Silverman's (1986) rule for the bandwidth of a kernel density estimator and Härdle and Marron's (1995) proposal for selection of the bandwidth of a kernel regression.

While zero level rules have the attraction of being simple and fast to compute, they also have the drawback of being heavily dependent on the choice of θ_{init} (see e.g., Jones, Marron and Sheather, 1996). They also lack theoretical properties such as consistency.

2.2. Level One Exact Risk Selectors

Level one exact risk selectors offer themselves as a remedy to the inadequacies of $\hat{h}_0$ by allowing the θ in the risk expression to be estimated nonparametrically. They are defined according to

$$\hat{h}_1 = \underset{h}{\operatorname{argmin}}\, R(h, \hat{\theta}_{\hat{g}})$$

where, for $g \in \mathcal{H}, \hat{\theta}_g$ is an estimate of θ based on the smoothing parameter g. The pilot smoothing parameter $\hat{g}$ must be chosen using

initial estimates that do not require specification of smoothing parameters.

Selection rules where $R(h, \theta)$ is replaced by $R(h, \hat{\theta}_g)$ for some g have been proposed by Müller (1985), Staniswalis (1989), Taylor (1989), Faraway and Jhun (1990), Hall, Marron and Park (1992) and Härdle, Hall and Marron (1992). The term 'double smoothing' is the most common name given to this type of strategy. In the kernel density estimation context it has also been referred to as smoothed cross-validation and bootstrap cross-validation, the later name due to the fact that it is equivalent to minimising the bootstrap risk in that setting.

In all of the above papers the choice of the pilot smoothing parameter is made using a simple rule, such as $g = h$, or through asymptotic rules such as minimising the estimated asymptotic variance of the selector. This raises the question of whether it is possible to instead select g using exact risk ideas.

In theory, one could aim to choose g to minimise $Q(g, \theta)$, where Q is some risk that measures the quality of

$$\hat{h}_g = \operatorname*{argmin}_h R(h, \hat{\theta}_g).$$

For $\mathcal{H} \subseteq \mathbb{R}$, the simplest such risk is the mean squared error of $\hat{h}_g$:

$$Q(g, \theta) = E_\theta\{(\hat{h}_g - h_\theta)^2\}.$$

Ideally our rule for choosing g would be an estimate of $g_\theta = \operatorname{argmin}_g Q(g, \theta)$. The main practical stumbling block to this approach is the fact that the discrepancy $\hat{h}_g - h_\theta$ is usually non-linear in the data and therefore its mean square does not have a closed form. Assuming that $\mathcal{H}$ is a continuous subset of $\mathbb{R}$, such as $(0, \infty)$, and that R is twice continuously differentiable with respect to h then formal Taylor series expansion of R about h_θ leads to the 'linearised' approximation:

$$\hat{h}_g - h_\theta \simeq \frac{-R^{[1]}(h_\theta, \hat{\theta}_g)}{R^{[2]}(h_\theta, \theta)} + \frac{R^{[1]}(h_\theta, \hat{\theta}_g)\{R^{[2]}(h_\theta, \hat{\theta}_g) - R^{[2]}(h_\theta, \theta)\}}{R^{[2]}(h_\theta, \theta)^2} \qquad (2)$$

where $R^{[i]}(h, \theta) = (\partial^i / \partial h^i) R(h, \theta)$. The advantage of this approximation is that it often does have a closed form mean square, particularly for

L_2-type risks. A convenient simplification is to ignore the second term in (2) in the hope that it has a lower order effect and work with

$$\operatorname*{argmin}_g E_\theta\left[\left\{\frac{-R^{[1]}(h_\theta, \hat{\theta}_g)}{R^{[2]}(h_\theta, \theta)}\right\}^2\right] = \operatorname*{argmin}_g E_\theta\{R^{[1]}(h_\theta, \hat{\theta}_g)^2\}.$$

This results in the rule

$$\hat{g} = \operatorname*{argmin}_g E_{\theta_{\text{init}}}\{R^{[1]}(h_{\theta_{\text{init}}}, \hat{\theta}_g)^2\}.$$

We call this approach *exact double smoothing*. The word 'exact' is used because none of the traditional 'smoothing parameter goes to zero' asymptotics are used to specify the rule.

If we define

$$L(g, h, \theta) = E_\theta\{R^{[1]}(h, \hat{\theta}_g)^2\}$$

then, clearly, $\hat{g}$ is the minimiser of $L(g, h_{\theta_{\text{init}}}, \theta_{\text{init}})$. In practice the remaining requirement for implementation of the rule is to find the expression for $L(g, h, \theta)$. In regression contexts with the MSSE risk this usually reduces to matrix calculus and determination of moments. For example, in the ordinary normal errors regression model (1) with the MSSE risk, standard results for moments of quadratic forms of a normal random vector lead to

$$\begin{aligned} L(g,h,m) =& \{m^T D_{gh} m + \sigma^2 \operatorname{tr}(D_{gh})\}^2 + 4\sigma^2 \operatorname{tr}(S_h' S_h^T) \\ & \times \{m^T D_{gh} m + \sigma^2 \operatorname{tr}(D_{gh})\} + 4\sigma^2 m^T D_{gh}^2 m + 2\sigma^4 \operatorname{tr}(D_{gh}^2) \\ & + 4\sigma^4 \operatorname{tr}^2(S_h' S_h^T) \end{aligned} \tag{3}$$

and

$$D_{gh} = S_g^T\{S_h'^T(S_h - I) + (S_h - I)^T S_h'\} S_g. \tag{4}$$

Here S_h' denotes the $n \times n$ matrix containing the derivatives of the entries of S_h with respect to h.

In summary, exact double smoothing rules for h take the form

$$\hat{h}_{\text{EDS}} = \operatorname*{argmin}_h R(h, \hat{\theta}_{\hat{g}}) \quad \text{where} \quad \hat{g} = \operatorname*{argmin}_g L(g, h_{\theta_{\text{init}}}, \theta_{\text{init}}). \tag{5}$$

3. FURTHER ILLUSTRATIONS

3.1. Kernel Density Estimation

Since the majority of the literature on second generational smoothing parameter selection has been in the kernel density estimation context, it is worth seeing where the exact risk approaches fit in. Based on a random sample $X_1, \ldots, X_n$ from a density f, the kernel density estimate of $f(x)$ is

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$$

where K is a kernel function, $h > 0$ is the bandwidth and $K_h(u) = K(u/h)/h$. The mean integrated squared error of $\hat{f}$ is

$$R(h, f) = E \int \{\hat{f}_h(x) - f(x)\}^2 dx$$

which, for symmetric K, can be shown to have the representation

$$R(h, f) = \int [\{(1 - n^{-1})K * K - 2K\}_h * f](x) f(x) dx.$$

Setting $\beta(u) = \{(1-n^{-1})\, K * K - 2K\}\,(u) + u\{(1-n^{-1})\, K' * K - 2K'\}(u)$, $D_{gh} = \beta_h * (K * K)_g$ and $\sigma_K^2 = \int K(u)^2 du$ we obtain

$$R^{[1]}(h, \hat{f}_g) = h^{-1} n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{gh}(X_i - X_j) - n^{-1} h^{-2} \sigma_K^2$$

which leads to

$$\begin{aligned} L(g, h, f) = & \left\{ n^{-1} D_{gh}(0) + (1 - n^{-1}) \int (D_{gh} * f)(x) f(x) dx \right\}^2 \\ & - 2n^{-1} h^{-2} \sigma_K^2 \left\{ n^{-1} D_{gh}(0) + (1 - n^{-1}) \right. \\ & \left. + (1 - n^{-1}) \int (D_{gh} * f)(x) f(x) dx \right\} \\ & + n^{-2} h^{-4} \sigma_K^4 + 2n^{-3}(n - 1) \left[\int (D_{gh}^2 * f)(x) f(x) dx \right. \\ & \left. - \left\{ \int (D_{gh} * f)(x) f(x) dx \right\} \right] \end{aligned}$$

$$+ 4n^{-3}(n-1)(n-2)\left[\int \{(D_{gh} * f)(x)\}^2 f(x)dx - \left\{\int (D_{gh} * f)(x) f(x)dx\right\}\right].$$

Standard Taylor series arguments can be used to show that, as $h \to 0$,

$$R^{[1]}(h, \hat{f}_g) \simeq h^3 \mu_2(K)^2 \int (\hat{f}_g'')^2 - n^{-1}h^{-2}\sigma_K^2$$
$$\simeq h^3 \mu_2(K)^2 \left\{\int (\hat{f}_g'')^2 - \int (f'')^2\right\} + R^{[1]}(h, f).$$

Therefore,

$$L(g, h_f, f) \simeq h_f^6 \mu_2(K)^2 E\left\{\int (\hat{f}_g'')^2 - \int (f'')^2\right\}^2.$$

so if f_{init} is a consistent approximation to f then the g that minimizes $L(g, h_{f_{\text{init}}}, f_{\text{init}})$ will be asymptotically the same as the g that minimises the mean squared error of $\int (\hat{f}_g'')^2$. It is known from, for example, Park and Marron (1992) that such a choice of g leads to optimal asymptotic performance of $\hat{h}_1$.

3.2. Smoothing Splines

For the model (1), the minimiser of the penalised least squares problem $\|Y - m\|^2 + h\int_0^1 (m'')^2$ can be shown to be the smoothing spline $\hat{m}_h = S_h Y$ where

$$S_h = (I + hK)^{-1} \tag{6}$$

and K is a symmetric matrix depending only on n (see e.g., Green and Silverman, 1994). In this case

$$S_h' = -(I - hK)^{-1} K (I + hK)^{-1} = h^{-1} S_h (S_h - I).$$

from which it can be easily shown that $D_{gh} = 2h^{-1} S_g S_h (S_h - I)^2 S_g$. Substitution into (3) and (5) yields the exact double smoothing rule for h.

We believe this to be the first 'second generational' rule for choosing the smoothing parameter of a smoothing spline. It would be interesting to study the theoretical and practical performance of $\hat{h}_1$ in this context and compare it with the popular GCV selection rule (Craven and Wahba, 1979).

3.3. Generalised Models

An important recent advance in smoothing technology is its extension to generalised linear model contexts (e.g., Hastie and Tibshirani, 1990, Green and Silverman, 1994). Suppose that the regression data (X_1, Y_1), ..., (X_n, Y_n) are modelled according to the quasi-likelihood model

$$E(Y_i|X_i) = g^{-1}\{\eta(X_i)\}$$
$$\text{var}(Y_i|X_i) = \sigma^2 V[g^{-1}\{\eta(X_i)\}]$$

where g is a link function and V is some positive function used to model the variance. For an arbitrary function f will use the notation $f(\eta) = [f\{\eta(X_1)\}, \ldots, f\{\eta(X_n)\}]^T$. Let

$$B_\eta = \text{diag}\{(g^{-1})'(\eta)\} \quad \text{and} \quad W_\eta = B_\eta^2\left(\text{diag}\ \{\sigma^2 V\{g^{-1}(\eta)\}\}\right)^{-1}.$$

and $\tilde{S}_\eta$ be the weighted least squares adaptation of the smoother matrix S with weights equal to the diagonal entries of W_η. For example, if S corresponds to the ordinary smoothing spline (6) then the W_η-weighted version is

$$\tilde{S}_\eta = (W_\eta + hK)^{-1}.$$

(Green and Silverman, 1994). Also define

$$\tilde{Y}_\eta = \eta + B_\eta^{-1}\{Y - g^{-1}(\eta)\},$$

sometimes called the adjusted dependent variable. Then, under appropriate conditions and a reasonable starting value $\hat{\eta}_0$, the sequence $\eta_1, \eta_2, \ldots$ converges to an estimate of η, $\hat{\eta}$, where

$$\hat{\eta}_{i+1} = \tilde{S}_{\eta_i}\tilde{Y}_{\eta_i}.$$

Most of the inference and diagnostics performed in generalised contexts is based on the 'one-step' approximation

$$\hat{\eta} \simeq \tilde{S}_\eta \tilde{Y}_\eta.$$

(e.g., Mc Cullagh and Nelder, 1988, Hastie and Tibshirani, 1990).

The same could be done for smoothing parameter selection by replacing S and Y in Section 2 by $\tilde{S}_\eta$ and Y_η, respectively. Noting that $\text{cov}(\tilde{Y}_\eta) = W_\eta^{-1}$, the risk would then be approximately

$$R(h, \eta) = \|(\tilde{S}_{\eta\,\text{init}} - I)\eta_{\,\text{init}}\|^2 + \text{tr}(\tilde{S}_{\eta\,\text{init}} W^{-1}_{\eta\,\text{init}} \tilde{S}^T_{\eta\,\text{init}})$$

and the exact double smoothing criterion function would be

$$\begin{aligned} L(g, h, \eta) = {} & \text{var}(\tilde{Y}_\eta^T \tilde{D}_{gh} \tilde{Y}_\eta) + \{\eta^T \tilde{D}_{gh} \eta + \text{tr}(\tilde{D}_{gh} W_\eta^{-1})\}^2 \\ & + 4\text{tr}(\tilde{S}'_{\eta\,\text{init}} W^{-1}_{\eta\,\text{init}} \tilde{S}^T_{\eta\,\text{init}})\{\eta^T \tilde{D}_{gh} \eta + \text{tr}(\tilde{D}_{gh} W_\eta^{-1})\} \\ & + 4\text{tr}^2(\tilde{S}'_{\eta\,\text{init}} W^{-1}_{\eta\,\text{init}} \tilde{S}^T_{\eta\,\text{init}}) \end{aligned}$$

where $\tilde{D}_{gh}$ defined similarly to D_{gh} in (4) but with $\tilde{S}_\eta$ rather than S. The variance expression depends on the third- and fourth-order moments of Y, and may require further modelling.

3.4. Partially Linear Models

The simplest partially linear model is

$$Y = m + \alpha + Z\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$$

where α is an $n \times 1$ constant vector, Z is an $n \times p$ design matrix, β is a $p \times 1$ vector of parameters and m is as in (1) but with the restriction $\Sigma\, m(x_i) = 0$ for identifiability purposes.

A consistent estimate of β is $\hat{\beta} = P_h Y$ where

$$P_h\{Z^T(I - S_h^*)Z\}^{-1} Z^T(I - S_h^*)$$

and $S_h^* = (I - n^{-1}J)S_h$ is the centered form of the smoother matrix. Here J is the $n \times n$ matrix of ones.

Consider the problem of estimating the parameter $\tau = c^T\beta$. Then, using the mean squared error as the risk, we obtain

$$R(h, m) = (c^T P_h m)^2 + \sigma^2 c^T P_h P_h^T c.$$

There are a few features about this smoothing parameter selection problem that distinguish it from the fully nonparametric case. Firstly, the asymptotically optimal smoothing parameter is a different order of magnitude. For example, if S_h corresponds to a local linear kernel estimator with bandwidth h then the optimal bandwidth is of order $n^{-1/3}$, compared to $n^{-1/5}$ for estimation of m (e.g., Opsomer & Ruppert, 1996). The other distinguishing feature is that there does not seem to be an analogue of cross-validation or Mallow' s C_L for this problem, so we are left with little choice but to seek a second generational approach if we are to choose the smoothing parameter optimally.

Applying the exact double smoothing principle one can show that the resulting rule for h is of the form given by (5) with

$$\begin{aligned} L(g, h, m) = &\{m^T D_{gh} m + \sigma^2 \mathrm{tr}(D_{gh})\}^2 \\ &+ 4\sigma^2 c^T P_h' P_h^T c\{m^T D_{gh} m + \sigma^2 \mathrm{tr}(D_{gh})\} \\ &+ 4\sigma^2 m^T D_{gh}^2 m + 2\sigma^4 \mathrm{tr}(D_{gh}^2) + 4\sigma^4 (c^T P_h' P_h^T c)^2 \end{aligned}$$

and $D_{gh} = S_g^{*T}(P_h'^T c c^T P_h + P_h^T c c^T P_h') S_g^*$.

3.5. Correlated Errors

Exact risk approaches to smoothing parameter selection are particularly attractive in the case of correlated errors since it is possible to account for the correlations through exact expressions. On the other hand, asymptotic approximations are often deficient in describing the effect of correlations (e.g., Hart, 1984 and Wand, 1992).

Consider the model

$$Y = m + \varepsilon, \quad \varepsilon \sim N(0, V)$$

where V is obtained by some correlation model for the ε_i's. Then the MSSE is

$$R(h, m) = \|(S_h - I)m\|^2 + \mathrm{tr}(S_h V S_h^T)$$

and the smoothing parameter can be selected using exact double smoothing with

$$L(g,h,m) = \{m^T D_{gh} m + \mathrm{tr}(D_{gh} V)\}^2 + 4\{m^T D_{gh} m + \sigma^2 \mathrm{tr}(D_{gh})\}\mathrm{tr}(S_h' V S_h^T) + 4m^T D_{gh} V D_{gh} m + 2\mathrm{tr}(D_{gh} V D_{gh} V) + 4\mathrm{tr}^2(S_h' V S_h^T)$$

and D_{gh} the same as in (4).

4. CHOOSING AN INITIAL ESTIMATE

Practical studies such as those in Park and Marron (1990), Sheather (1992), Cao, Gonzalez-Manteiga and Ceuvas (1994) and Jones, Marron and Sheather (1995) have shown that second generation smoothing parameter selectors possess a reasonably high degree of robustness against misspecification of the initial estimate. Nevertheless, as pointed out by Janssen, Marron, Veraverbeke and Sarle (1995) and Loader (1995), it is always possible for a smoothing parameter selector to perform arbitrarily poorly by having the true curve sufficiently different from the class of possible initial estimates. The same is, of course, true for the exact risk selectors described in the previous section, including exact double smoothing. The most appropriate choice of the initial estimate depends on many factors such as the type of functions that usually arise in the application area (e.g., biomedical applications tend to involve much smoother and less detailed regression functions than those typically arising in electrical engineering) and the level of complexity which the user is willing to deal with.

If the user places a high premium on simplicity then suitable initial estimators exist in the literature. For the nonparametric regression problem Härdle and Marron (1995) developed a class of simple initial estimators based on blockwise polynomial fitting. Ruppert, Sheather and Wand (1995) extended this idea by suggesting that the number of blocks be chosen by Mallows' criterion. In density estimation the simplest initial estimate is the normal reference rule (e.g., Scott 1979) mentioned in Section 2.1. More sophisticated initial density estimates, which aim to correct the shortcomings of the normal reference rule, include the skewness and kurtosis adjusted estimates of Scott (1992, pp. 56–57) and the scale measure estimates of Janssen *et al.* (1995).

Another new path that one might consider, at least for the ordinary nonparametric regression problem, is to take the initial estimate to be a Bayesian regression spline smoother of the type developed by Smith and Kohn (1996). Apart from being very flexible, these have the appealing feature of being, in some sense, smoothing parameter free. The reason for this is that the knot selection is handled through a Bayesian approach with the estimator appearing to be quite insensitive to the choice of the prior. An S-PLUS/Fortran module that facilitates the computation of a Bayesian regression spline smooth, written by M. Smith, is available at the World Wide Web site http://lib.stat.cmu.edu/S/br. Simulation studies given in Section 5 indicate that this approach results in a very effective bandwidth selector for the local linear kernel estimator. For the density estimation problem a rough analogue of the Bayessian regression spline smoother that might be worth considering is the Bayesian normal mixture density estimator of Roeder and Wasserman (1995). However, we have not yet investigated this proposal.

An obvious question that might be raised about the Bayesian regression spline approach is: if one has such a good smoothing parameter-free smoother at one's disposal then why not simply use this as the actual smoother? Our response is that it depends on the user's personal tastes. Many analysts like the simplicity of the local polynomial smoother since it is easy to see how it uses the data. On the other hand, Bayesian regression splines are more of a 'black box' to most practitioners. A more extreme example is the histogram, which is used by almost every expert and non-expert data analyst. We see no conflict with providing these users with a high performance automatic smoothing parameter choice, even if it is based on more complex methodology (see e.g., Wand, 1997).

5. SIMULATION RESULTS

A simulation study was run to compare the exact double smoothing smoothing parameter selector to some other approaches. The setting was ordinary nonparametric regression

$$Y_i = m\{(i-1)/99\} + \sigma\varepsilon_i, \quad i = 1, \ldots, 100$$

with the ε_i being independent standard normal random variables. The following functions and corresponding noise levels were considered:

$$\text{Exponential: } m(x) = e^{-5x}, \qquad \sigma = 0.1$$
$$\text{Sinusoidal: } m(x) = \sin(5\pi x), \qquad \sigma = 0.5$$

and the number of replications was 200. The cross-validatory selector and the direct plug-in selector of Ruppert, Sheather and Wand (1995) were also computed for comparison. The exact double smoothing selector used the Smith & Kohn (1996) Bayesian regression spline smoother as an initial estimate of m. An estimate of σ^2 was obtained by dividing the residual sum of squares of this initial estimate by the sample size minus the number of fitted parameters.

Figure 1 summarises the results through histograms of values of $\log(\hat{h}/h_m)$. A tight distribution around zero can be equated with good performance of the selector in terms of estimation of h_m. We see that the exact double smoothing selector does well for both curves, being more stable than cross-validation and direct plug- in. In the case of the

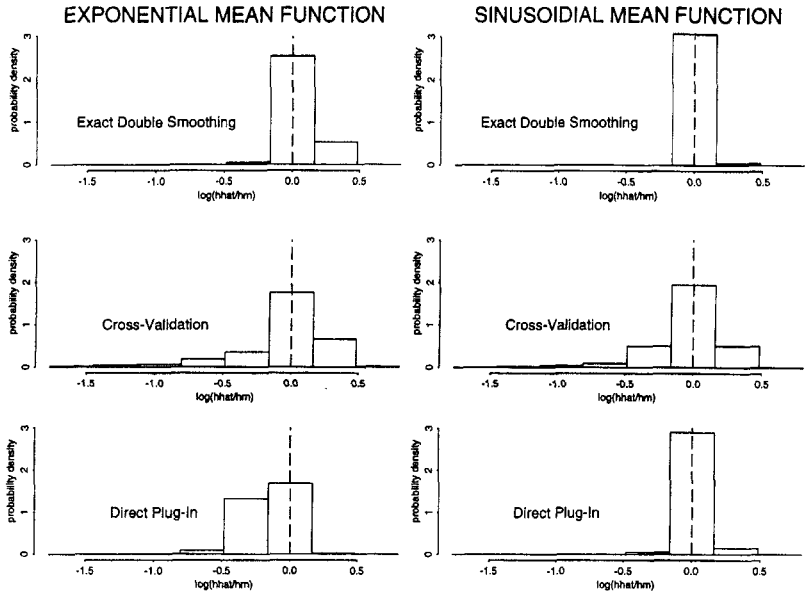

FIGURE 1 Histograms of $\log(\hat{h}/h_m)$ from the simulation study described in the text.

exponential mean function direct plug-in appears to be biased to the left due to the large boundary effects which asymptotic approximations do not describe very well. Exact double smoothing appears to correct this problem.

6. CONCLUSION

We have shown that exact risk approaches to smoothing parameter selection are simpler to specify and more versatile than existing second-generational approaches. They are also less dependent on asymptotic approximations which often leads to improved performance. Our simulations have indicated that, when combined with a good initial estimate such as the Bayesian regression spline of Smith & Kohn, exact double smoothing results in a very good smoothing parameter selector for local linear regression. The heuristics given at the end of Section 3.1 suggest that exact double smoothing also has good theoretical properties, although this warrants further investigation.

Acknowledgements

This work was partially supported by grants from the Australian Research Council. Programming assistance from Ilse Augustyns and Paul Yau is gratefully acknowledged. The first author is thankful to the Institute of Mathematical Statistics, Lund Technical University, Sweden and to the Department of Statistics, University of Illinois, U.S.A., for hospitality provided during the course of this research. This research was completed in part while the second author was at the Australian Graduate School of Management, University of New South Wales, Australia.

References

Cao, R., Cuevas, A. and González-Manteiga, W. (1994) A comparative study of several smoothing methods in density estimation. *Comp. Statist. Data Anal.*, **17**, 153–76.

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions. *Numer. Math.*, **31**, 377–403.

Fan, J. and Gijbels, I. (1995) Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society B*, **57**, 371–394.

Faraway, J. J. and Jhun, M. (1990) Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.*, **85**, 1119–1122.

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

Hall, P., Marron, J. S. and Park, B. U. (1992) Smoothed cross-validation. *Probab. Theory Rel. Fields*, **92**, 1–20.

Hall, P., Marron, J. S. and Titterington, D. M. (1995) On partially local smoothing rules for curve estimation. *Biometrika*, **82**, 575–588.

Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimum *J. Amer. Statist. Assoc.*, **83**, 86–95.

Härdle, W., Hall, P. and Marron, J. S. (1992) Regression smoothing parameters that are not far from their optimum. *J. Amer. Statist. Assoc.*, **87**, 227–33.

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

Hart, J. D. (1984) Efficiency of a kernel density estimator under an autoregressive dependence model. *J. Amer. Statist. Assoc.*, **79**, 110–117.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman and Hall, London.

Janssen, P., Marron, J. S., Veraverbeke, N. and Sarle, W. (1995) Scale measures for band-selection. *J. Nonparametric Statist.*, **5**, 359–80.

Jones, M. C. and Kappenman, R. F. (1992) On a class of kernel density estimate bandwidth selectors, *Scand. J. Statist.*, **19**, 337–50.

Jones, M. C., Marron, J. S. and Park, B. U. (1991) A simple root-n bandwidth selector. *Ann. Statist.*, **19**, 1919–32.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996) A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.*, **19**, 401–407.

Kim, W. C., Park, B. U. and Marron, J. S. (1994) Asymptotically best bandwidth selectors in kernel density estimation. *Statist. Probab. Lett.*, **19**, 119–27.

Loader, C. R. (1995) Old Faithful erupts: bandwith selection reviewed, unpublished manuscript.

Mallows, C. L. (1973) Some comments on C_p. *Technometrics*, **15**, 661–675.

McCullagh, P. and Nelder, J. A. (1988) *Generalized Linear Models*. Second Edition. Chapman and Hall, London.

Müller, H.-G. (1985) Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. *Statist. Decisions* Supplement no. **2**, 193–206.

Opsomer, J. D. and Ruppert, D. (1996) A root-n consistent backfitting estimator for semiparametric additive modelling. Unpublished manuscript.

Park, B. U. and Marron, J. S. (1990) Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.*, **85**, 66–72.

Park, B. U. and Marron, J. S. (1992) On the use of pilot estimators in bandwidth selection. *J. Nonparametric Statist.*, **1**, 231–40.

Roeder, K. and Wasserman, L. (1995) Practical Bayesian density estimation using mixtures of normals. Carnegie Mellon University, Department of Statistics, Tech. Report, No. 633.

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

Scott, D. W. (1979) On optimal and data-based histograms. *Biometrika* **66**, 605–10.

Scott, D. W. (1985) Frequency polygons: theory and applications. *J. Amer. Statist. Assoc.*, **80**, 348–354.

Scott, D. W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York.

Sheather, S. J. (1992) The performance of six popular bandwidth selection methods on some real data sets (with discussion). *Comput. Statist.*, **7**, 225–50, 271–281.

Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser.*, **B 53**, 683–690.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection, *J. Econometrics*, **75**, 317–344.

Staniswalis, J. G. (1989) Local bandwidth selection for kernel estimates. *J. Amer. Statist. Assoc.*, **84**, 284–288.

Taylor, C. C. (1989) Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika*, **76**, 705–712.

Wand, M. P. (1992) Finite sample performance of density estimators under moving average dependence. *Statist. Probab. Lett.*, **13**, 109–15.

Wand, M. P. (1997) Data-based choice of histogram binwidth, *The American Statistician*, **51**, 59–64.