Comput Stat
DOI 10.1007/s00180-015-0577-7

ORIGINAL PAPER



# Practical use of robust GCV and modified GCV for spline smoothing

**Mark A. Lukas**[1] · **Frank R. de Hoog**[2] · **Robert S. Anderssen**[2]

Received: 25 July 2014 / Accepted: 25 March 2015


**Abstract** Generalized cross-validation (GCV) is a popular parameter selection criterion for spline smoothing of noisy data, but it sometimes yields a severely undersmoothed estimate, especially if the sample size is small. Robust GCV (RGCV) and modified GCV are stable extensions of GCV, with the degree of stabilization depending on a parameter $\gamma \in (0, 1)$ for RGCV and on a parameter $\rho > 1$ for modified GCV. While there are favorable asymptotic results about the performance of RGCV and modified GCV, little is known for finite samples. In a large simulation study with cubic splines, we investigate the behavior of the optimal values of $\gamma$ and $\rho$, and identify simple practical rules to choose them that are close to optimal. With these rules, both RGCV and modified GCV perform significantly better than GCV. The performance is defined in terms of the Sobolev error, which is shown by example to be more consistent with a visual assessment of the fit than the prediction error (average squared error). The results are consistent with known asymptotic results.

**Keywords** Generalized cross-validation · Prediction error · Sobolev error · Smoothing parameter · Spline smoothing

✉ Mark A. Lukas
M.Lukas@murdoch.edu.au

Frank R. de Hoog
Frank.deHoog@csiro.au

Robert S. Anderssen
Bob.Anderssen@csiro.au

1 Mathematics and Statistics, Murdoch University, South Street, Murdoch, WA 6150, Australia

2 CSIRO Mathematics, Informatics and Statistics, GPO Box 664, Canberra, ACT 2601, Australia





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

## 1 Introduction

The smoothing of observational data is a common and important problem in applications. We will assume that the data can be modeled as

$$y_i = f(x_i) + \varepsilon_i, \quad a \le x_1 < x_2 < \cdots < x_n \le b, \quad i = 1, \ldots, n, \tag{1}$$

where $f(x)$ is a smooth function and the random errors $\varepsilon_i$ are independent with mean 0 and common variance $\sigma^2$. A popular approach for estimating $f(x)$ is to use a smoothing spline (Eubank 1988; Green and Silverman 1994; Gu 2002; Wahba 1990).

The natural polynomial smoothing spline of degree $2m-1$ is defined as the minimizer $f_\lambda$ of

$$n^{-1}\sum_{i=1}^{n}(y_i - h(x_i))^2 + \lambda\int_a^b (h^{(m)}(x))^2 dx \tag{2}$$

over all functions $h$ for which $h^{(m)}$ is square integrable. The most common case is when $m = 2$, which yields the cubic smoothing spline. The amount of smoothing is determined by the smoothing parameter $\lambda$, and its choice is crucial to obtaining a good estimate $f_\lambda$.

One of the most popular parameter selection criteria is generalized cross-validation (GCV) (Craven and Wahba 1979; Wahba 1990). Let $A(\lambda)$ be the smoothing matrix defined by $\boldsymbol{f}_\lambda = A(\lambda)\boldsymbol{y}$, where $\boldsymbol{f}_\lambda = (f_\lambda(x_1), \ldots, f_\lambda(x_n))^T$. The GCV criterion selects $\lambda$ as the minimizer of the GCV function

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

where $\|\cdot\|$ is the Euclidean norm.

Let $T(\lambda)$ denote the (mean square) prediction error $T(\lambda) = n^{-1}\|\boldsymbol{f} - \boldsymbol{f}_\lambda\|^2$, where $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))^T$. It is known that the GCV function (up to an additive constant) behaves like an unbiased estimate of the prediction risk $ET(\lambda) = n^{-1}E\|\boldsymbol{f} - \boldsymbol{f}_\lambda\|^2$, and the criterion is asymptotically optimal with respect to the prediction error and risk as the sample size $n \to \infty$ (Cox 1984; Craven and Wahba 1979; Gu 2002; Li 1986; Nychka 1990). However, for small or moderately sized samples, GCV can be unstable, leading to severely undersmoothed estimates for a significant proportion of cases (Efron 2001; Wahba 1990, Sect. 4.9). This is evident graphically in that the GCV function can be very flat near 0 and may have more than one local minimum.

Motivated by this deficiency, a more stable extension of GCV, called robust GCV (RGCV), has been proposed (Lukas 2006, 2008; Robinson and Moyeed 1989). The RGCV criterion selects $\lambda$ as the minimizer of

$$\overline{V}(\lambda) = [\gamma + (1-\gamma)\mu_2(\lambda)]V(\lambda), \tag{4}$$

where $\mu_2(\lambda) = n^{-1}\mathrm{tr}(A^2(\lambda))$ and $\gamma \in (0, 1)$ is a robustness parameter. It has been shown in de Hoog et al. (2011), Lukas et al. (2010) that the RGCV function $\overline{V}(\lambda)$ in (4), like the GCV function in (3), can be computed efficiently in $O(m^2 n)$ operations.





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

It is known (Wahba 1990) that $\mu_2(0) = 1$ and $\mu_2(\lambda)$ is a monotonic decreasing function that approaches $m/n$ as $\lambda \to \infty$. Therefore, letting $\delta = \gamma + (1-\gamma)(m/n) \in (m/n, 1)$, we have $(1/\delta)\overline{V}(\lambda) \sim V(\lambda)$ as $\lambda \to \infty$, while $(1/\delta)\overline{V}(\lambda) \sim (1/\delta)V(\lambda) > V(\lambda)$ as $\lambda \to 0$. This shows that, compared to $V(\lambda)$, the graph of $\overline{V}(\lambda)$ has a modified shape near 0, which makes RGCV less likely than GCV to select an extremely small value of $\lambda$ (Lukas 2006, 2008).

The RGCV criterion for spline smoothing has been analyzed in Lukas et al. (2012). It is shown there, using a geometric approach developed in Efron (2001), that, as $\gamma$ is decreased from 1 (the GCV case), the stability of RGCV improves considerably, and it is very stable when $\gamma = 0.3$.

Another approach to stabilizing GCV is the modified GCV criterion (Cummins et al. 2001; Kim and Gu 2004), defined by the score function

$$V_\rho(\lambda) = \frac{n^{-1}\|(I - A(\lambda))\boldsymbol{y}\|^2}{[n^{-1}\mathrm{tr}(I - \rho A(\lambda))]^2}, \tag{5}$$

where $\rho > 1$ is a stabilization parameter. When $\rho = 1$, the method is just GCV, and, as $\rho$ is increased, the method becomes more stable. The effect of $\rho$ is to modify the shape of $V(\lambda)$ and to constrain the effective degrees of freedom $\mathrm{tr}A(\lambda)$ to be less than $n/\rho$. It is known (Cummins et al. 2001; Lukas 2008) that, under mild conditions, for cubic spline smoothing, modified GCV has the same asymptotic behavior as RGCV as $n \to \infty$, provided that the respective parameters $\rho$ and $\gamma$ are related by the formula

$$\gamma^{-1} = 1 + 8(\rho - 1)/3. \tag{6}$$

In almost all of the existing literature on spline smoothing, the prediction error (or corresponding integrated squared error) is the only performance measure considered. However, it is not an ideal measure because it is insensitive to deviations in the derivative and curvature of a spline estimate $f_\lambda$, deviations that can easily be detected by eye. As a consequence, the prediction error can be very misleading by failing to discriminate extreme undersmoothing, especially when the sample size is small (Lukas 2014).

In a visual assessment of the accuracy of a fitted curve (when compared to the underlying function), one intuitively takes into account not only function values but also the first derivative and second derivative, i.e. the linearized curvature, (and possibly even higher derivatives) of the curve. Assume that $f$ has derivatives of order $1, \ldots, m$, and $f^{(m)}$ belongs to $L_2$. Then, since $f_\lambda^{(m)}$ belongs to $L_2$ by definition, both $f_\lambda$ and $f$ belong to the Sobolev space $\mathcal{W} = W^{m,2}[a, b]$. Hence, an appropriate measure of accuracy is the (squared) Sobolev error (Lukas et al. 2012)

$$W(\lambda) := \|f - f_\lambda\|_{\mathcal{W}}^2 = \int_a^b (f(x) - f_\lambda(x))^2 dG + \kappa \int_a^b (f^{(m)}(x) - f_\lambda^{(m)}(x))^2 dx, \tag{7}$$

where $\kappa = (b-a)^{2m-1}$ and $G$ (assumed differentiable) is the limit as $n \to \infty$ of the empirical distribution functions for the design points $x_i$. Note that, although only the



Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

$m$th derivative appears in the Sobolev error, it is automatically sensitive to deviations in derivatives of lower order (Lukas et al. 2012).

The weight $\kappa = (b-a)^{2m-1}$ is used in (7) to standardize the form of the error so that it is independent of the scale of the interval $[a, b]$. To see this, let $h(x)$ be any smooth function of $x \in [a, b]$, and let $\bar{h}(t) = h(a + t(b - a))$ and $\bar{G}(t) = G(a + t(b - a))$, where $t \in [0, 1]$, be the corresponding functions scaled onto $[0, 1]$. The first term on the right-hand side of (7) is independent of the scale because

$$\int_a^b (h(x))^2 \frac{dG}{dx}\, dx = \int_0^1 (\bar{h}(t))^2 \frac{d\bar{G}}{dt}\, dt.$$

For the second term, since $d^m h/dx^m = (b-a)^{-m} d^m \bar{h}/dt^m$, we have

$$(b-a)^{2m-1} \int_a^b (h^{(m)}(x))^2 dx = \int_0^1 (\bar{h}^{(m)}(t))^2 dt.$$

Hence (with $h = f - f_\lambda$), the Sobolev error $W(\lambda)$ is independent of the length of the interval $[a, b]$.

Clearly, any weighted Sobolev error of the form

$$W_{\mathbf{c}}(\lambda) := c_1 \int_a^b (f(x) - f_\lambda(x))^2 dG + c_2 \kappa \int_a^b (f^{(m)}(x) - f_\lambda^{(m)}(x))^2 dx, \qquad (8)$$

where $c_1 > 0$ and $c_2 > 0$ are independent of $a$ and $b$, is also invariant under horizontal scaling. The first integral in (8) is the integrated squared error corresponding to the prediction error, so the ratio $c_2/c_1$ determines the relative importance placed on the error in $f^{(m)}$ compared to the prediction error. The choice of $c_1$ and $c_2$ is a subjective one, but, obviously, the simplest choice is $c_1 = c_2 = 1$ as in (7). Furthermore, as will be seen in Sect. 3, this choice is also consistent with a visual assessment of different spline fits for our first example. It is shown in Lukas (2014), by asymptotic analysis and simulations, that the Sobolev error $W(\lambda)$ is significantly better than the prediction error in discriminating extreme undersmoothing.

The asymptotic performance of RGCV and of modified GCV for spline smoothing, with respect to both the prediction risk and Sobolev risk, have been analyzed in Lukas et al. (2012). For cubic splines, it was shown that, under suitable assumptions, the asymptotic efficiency for the Sobolev risk of the (restricted) RGCV estimator improves as $\gamma$ is decreased from 1 (the GCV case) until it reaches an extremum at a unique value of $\gamma \in (0, 0.6)$. A similar result was also shown for modified GCV.

This paper investigates the finite sample behavior of RGCV and modified GCV, through a large simulation study for cubic spline smoothing. The aim is to determine the effect of the parameters $\gamma$ and $\rho$ on the performance of the methods, and to give simple practical rules for choosing these parameters. Robinson and Moyeed (1989) state that $\gamma = 0.5$ worked well in their simulations with cubic splines and prediction error loss, but no results for other values of $\gamma$ were presented. Based on simulations with cubic splines and prediction error loss, Cummins et al. (2001) and Kim and Gu





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

(2004) suggest that $\rho = 1.4$ is a good choice for modified GCV. We obtain more detailed results about the optimal selection of these parameters.

The paper is organized as follows. After discussing computational issues in Sect. 2, we present simulation results in Sect. 3 to demonstrate the enhanced performance of RGCV compared to GCV, and the effectiveness of the Sobolev error over the prediction error as a measure of performance. Section 4 contains the results of simulations of the RGCV criterion with the Sobolev error as the performance measure. We identify the following simple rule for choosing the parameter $\gamma$:

$$\gamma_{\text{rule}} = \begin{cases} 0.2, & \text{if } n < 100, \\ 0.3, & \text{if } n \geq 100, \end{cases}$$

and show that, for the examples considered, this rule is close to optimal in minimizing the mean inefficiency for the Sobolev error.

Section 5 describes the results of the corresponding simulations for the modified GCV criterion. We identify the following simple rule for choosing the parameter $\rho$:

$$\rho_{\text{rule}} = \begin{cases} 1.3, & \text{if } n < 100, \\ 2, & \text{if } n \geq 100, \end{cases}$$

and show that it is close to optimal. In Sect. 6, the RGCV results are compared directly with those for modified GCV. With optimal parameter values, RGCV and modified GCV have very similar performance for large $n$ (consistent with the asymptotic theory), but modified GCV appears to have the better performance for small $n$. With the values $\gamma_{\text{rule}}$ and $\rho_{\text{rule}}$ above, RGCV and modified GCV have comparable performance.

It is of interest to know how the near-optimal rules above are affected by different choices of $c_1$ and $c_2$ if the weighted Sobolev error $W_{\mathbf{c}}(\lambda)$ in (8) is used as the performance measure. First, it is shown in Lukas et al. (2012) that, if the function $f$ has smoothness up to a certain level (but not if $f$ is "very smooth" (Nychka 1990), i.e. not if $f$ is in $W^{2m,2}[a, b]$ and satisfies the natural boundary conditions $f^{(j)}(a) = f^{(j)}(b) = 0$, $j = m, \ldots, 2m - 1$), the asymptotically optimal (with respect to $W_{\mathbf{c}}(\lambda)$) choices of $\gamma$ for RGCV and of $\rho$ for modified GCV are independent of $c_1$ and $c_2$. This means that the optimal values of $\gamma$ and $\rho$ for $W_{\mathbf{c}}(\lambda)$ will plateau as $n$ becomes large, and so, for $m = 2$, they may not change very much from the near-optimal values $\gamma = 0.3$ and $\rho = 2$ found above for $n$ as high as 1600. However, for small or moderately-sized values of $n$, and also for large $n$ if $f$ is "very smooth", we can expect the choice of $c_2/c_1$ to have a much greater effect. In particular, if $c_2/c_1$ is decreased (from 1) and becomes close to 0, then $W_{\mathbf{c}}(\lambda)$ behaves like the prediction error, so the optimal value of $\gamma$ will increase (from $\gamma_{\text{rule}}$) and the optimal value of $\rho$ will decrease (from $\rho_{\text{rule}}$). These trends follow for finite $n$ because the asymptotically optimal values of $\gamma$ and $\rho$ for the prediction error are both equal to 1 (as GCV is asymptotically optimal for the prediction error).

Recently, Girard (2010, Sect.7.2) developed a novel approach to modify GCV to provide more smoothing when it is needed. In this approach, the GCV parameter choice is perturbed by a positive quantity such that the resulting parameter choice is the $(1 - \alpha)100$th percentile of the randomisation-based distribution of $\hat{\lambda}_{RandGCV}$,





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

obtained by replacing the trace term in (3) with a randomised version that uses a simulated white noise vector. Asympotic results in Girard (2010) yield the simple bound $1 + z_{1-\alpha}^2$, where $z_{1-\alpha}$ is the $(1-\alpha)100$th percentile of $Z \sim N(0, 1)$, for the relative asymptotic increase of the prediction error loss for the new choice (relative to GCV) on average, also called the relative risk regret. An accurate determination of the $(1-\alpha)100$th percentile of $\hat{\lambda}_{RandGCV}$ is computationally expensive, so RGCV and modified GCV have a computational advantage for spline smoothing since their parameter choices can be computed efficiently. It would be interesting to compute and compare the randomisation-based approach with RGCV and modified GCV in simulations with our examples and the Sobolev error, but it is not done in this paper.

## 2 Computational issues

For our simulations, we consider five test functions with synthetic data $y_i = f(x_i) + \varepsilon_i$, $i = 1, \ldots, n$, where $x_i = a + (b-a)(i-1)/(n-1)$ are equally spaced on $[a, b]$ and $\varepsilon_i$ are pseudo-random i.i.d. normal variates from $N(0, \sigma^2)$, for three different values of $\sigma$. The sample size $n$ ranges from 20 to 1600. The first three functions are Examples 1–3 from Craven and Wahba (1979), in which

$$f(x) = \sum_{j=1}^{r} w_j \beta_{p_j, q_j}(x), \quad x \in [0, 1], \tag{9}$$

where $\beta_{pq}(x)$ is the beta density function

$$\beta_{pq}(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1}$$

and the parameter values are

$$\begin{array}{lll} \text{Example 1}: r = 3 & \mathbf{w} = [0.2, 0.7, 0.1] & \mathbf{p} = [4, 5, 12] \quad \mathbf{q} = [15, 7, 5] \\ \text{Example 2}: r = 2 & \mathbf{w} = [0.4, 0.6] & \mathbf{p} = [12, 4] \quad \mathbf{q} = [7, 11] \\ \text{Example 3}: r = 3 & \mathbf{w} = [0.5, 0.2, 0.3] & \mathbf{p} = [10, 20, 30] \quad \mathbf{q} = [30, 20, 10]. \end{array}$$

These functions are unimodal, bimodal and trimodal, respectively, and their graphs are shown in Figs. 1a and 5a, b, respectively.

We also consider the example of the Motorcycle function (Cummins et al. 2001) defined by

$$f(x) = I_{x>0} \sin(2\pi(1-x)^2), \quad x \in [-0.2, 1], \tag{10}$$

where $I_{x>0}$ is 1 for $x > 0$ and 0 otherwise. This test example is difficult because $f'(x)$ is discontinuous at $x = 0$ (see Fig. 12a). The last example is the function

$$f(x) = -\sin(x/(0.1 + 0.4(x-0.5)^2)), \quad x \in [0, 1], \tag{11}$$

which we call a 'peaks' function. This yields a difficult smoothing problem because of the close proximity of two similar peaks (see Fig. 13a).





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

It is well known (Anselone and Laurent 1968; Hutchinson and de Hoog 1985; Reinsch 1967, 1971) that the polynomial spline $f_\lambda$ minimizing (2) can be computed by expressing $f_\lambda^{(m)}$ as a linear combination of B-splines. The coefficients are defined by a linear system with a symmetric positive definite banded matrix $P = P(\lambda)$ of bandwidth $2m+1$, and they can be found efficiently using a banded Cholesky decomposition. We use the representation of the cubic spline $f_\lambda$ described in Reinsch (1967).

The main difficulty in computing the GCV score function (3) is the evaluation of $\mathrm{tr}A$. This can be computed efficiently using the algorithm in Hutchinson and de Hoog (1985), which involves $O(m^2 n)$ operations. For the RGCV score function (4), in addition to $V(\lambda)$, one must calculate $\mathrm{tr}A^2$. Recently, two new efficient $O(m^2 n)$ algorithms were developed to calculate both $\mathrm{tr}A$ and $\mathrm{tr}A^2$ (de Hoog et al. 2011; Lukas et al. 2010).

One of these algorithms uses the expressions

$$\lambda^{-1}\mathrm{tr}(I - A(\lambda)) = \frac{d}{d\lambda}\log(\det P(\lambda)),$$

$$\lambda^{-2}\mathrm{tr}(I - A(\lambda))^2 = -\frac{d^2}{d\lambda^2}\log(\det P(\lambda)),$$

along with an efficient method for computing the derivatives of the log determinant, utilizing the band structure of $P(\lambda)$. This is the algorithm used in the simulations for this article. For cubic splines, the algorithm requires approximately $43n$ operations to compute both $\mathrm{tr}(I - A)$ and $\mathrm{tr}A^2$ (Lukas et al. 2010). This good computational efficiency allows us to carry out the extensive simulations needed to determine the behavior of RGCV for $n$ as large as 1600.

The simulations were implemented in MATLAB. Where appropriate, the powerful sparse matrix capabilities of MATLAB were utilized to improve the speed of computation. The GCV, RGCV and modified GCV estimates were computed by a simple minimization of $V(\lambda)$, $\overline{V}(\lambda)$ and $V_\rho(\lambda)$, respectively, on the grid with $\log_{10}(\lambda)$ ranging from $-12$ to $-1$ with spacing 0.2. This gave a good balance between speed and accuracy.

The prediction error depends only on function values at the points $x_i$ and so it is easily calculated after the spline $f_\lambda$ has been computed. The Sobolev error (7) is more difficult to calculate because of the derivatives and integrals involved. Since the points $x_i$ are equally spaced, we have $dG = (b-a)^{-1}dx$ in (7).

For simplicity, we chose to calculate the Sobolev error using the approximation

$$W(\lambda) = \|f - f_\lambda\|_{\mathcal{W}}^2 \approx \|f_{\mathrm{int}} - f_\lambda\|_{\mathcal{W}}^2, \tag{12}$$

where $f_{\mathrm{int}}$ is the natural cubic spline interpolating $f(x)$ at $x_1, \ldots, x_n$. The spline $f_{\mathrm{int}}$ is computed in a similar way to $f_\lambda$, and the right-hand side of (12) is easy to calculate using (7), because $f_{\mathrm{int}}$ and $f_\lambda$ are both splines with the same knots.

Because each function $f$ in our simulations is smooth (piecewise smooth in one example), it is approximated well by $f_{\mathrm{int}}$. The error $\|f - f_{\mathrm{int}}\|_{\mathcal{W}}$ is relatively small compared to $\|f - f_\lambda\|_{\mathcal{W}}$, so the approximation on the right-hand side of (12) is quite accurate, even for $n$ as small as 20. Using the same argument as for Lemma 1 in Lukas





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

et al. (2012), it is clear that $f_{\text{int}}$ is actually the best approximation to $f$ in the space of natural spline functions with knots at $x_i$, with respect to the squared norm defined by

$$\|h\|^2_{\widetilde{\mathcal{W}}} := n^{-1}\sum_{i=1}^{n}(h(x_i))^2 + (b-a)^{2m-1}\int_a^b (h^{(m)}(x))^2 dx, \tag{13}$$

and, consequently, $\|f - f_{\text{int}}\|_{\widetilde{\mathcal{W}}} \le \|f - f_\lambda\|_{\widetilde{\mathcal{W}}}$ for all $\lambda$. Note that (13) is a partially discretized form of the squared Sobolev norm $\|h\|^2_{\mathcal{W}}$.

For each smoothing parameter estimate $\hat{\lambda}$ computed, we can calculate the inefficiency for the prediction error, i.e. $I_T(\hat{\lambda}) = T(\hat{\lambda})/\min T(\lambda)$, and the inefficiency for the Sobolev error, i.e. $I_W(\hat{\lambda}) = W(\hat{\lambda})/\min W(\lambda)$. The denominators were computed by minimizing $T(\lambda)$ and $W(\lambda)$ on the grid with $\log_{10}(\lambda)$ ranging from $-15$ to 0 with spacing 0.2.

## 3 Comparison of GCV and RGCV

In this section, we will use both the prediction error $T(\lambda)$ and the Sobolev error $W(\lambda)$ in simulations to compare the performance of the GCV and RGCV criteria. The problem considered is Ex. 1 in (9) with $\sigma = 0.1$ and small sample size of $n = 20$. The function $f(x)$ and two different data sets are shown in Fig. 1a, b.

To illustrate the instability and undersmoothing behavior of the GCV criterion for this problem, Fig. 2a shows 20 replicates of the GCV function (3), with the corresponding GCV estimates (i.e. the minimizers of $V(\lambda)$) marked with a + symbol under the graphs. Clearly, there is a lot of variability in the function replicates and the estimates, consistent with the instability results in Efron (2001). Figure 2b shows the corresponding 20 replicates of the RGCV function (4) with $\gamma = 0.1$, together with the RGCV estimates. Clearly, compared to the GCV function replicates, the RGCV function replicates have higher curvature near their minimum points, and so their minimizers are well defined. Consequently, the RGCV estimates are much more stable with no undersmoothing behavior, consistent with the stability results in Lukas et al. (2012).

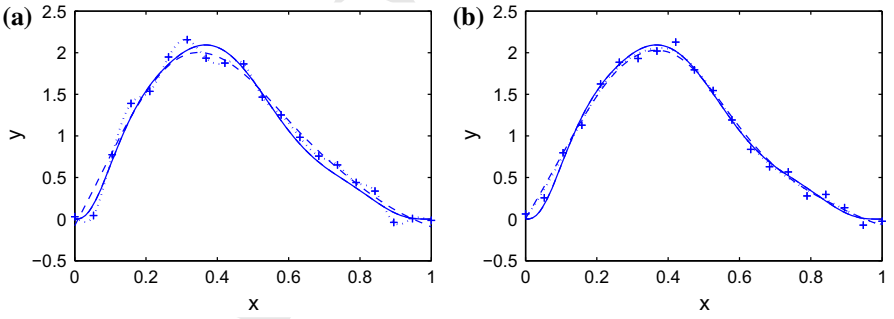


**Fig. 1** Function $f(x)$ (*solid*), data (+), GCV spline (*dotted*) and RGCV ($\gamma = 0.1$) spline (*dashed*) for **a** replicate 2 and **b** replicate 5 for Ex. 1 with $\sigma = 0.1$ and $n = 20$



Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

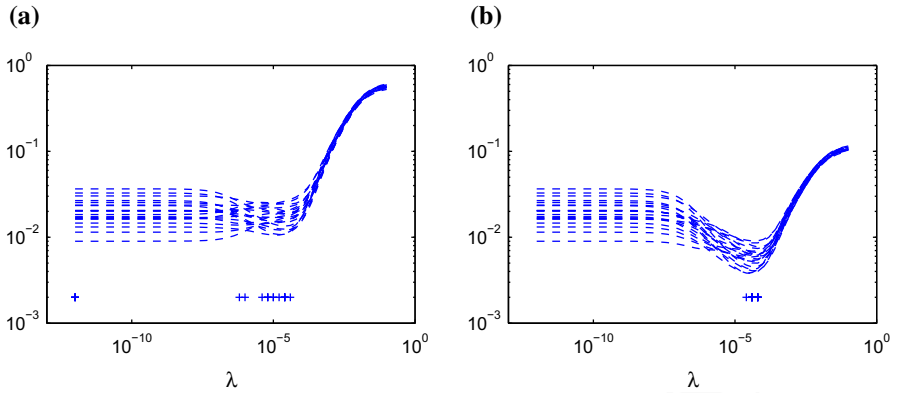


**Fig. 2** Twenty replicates of **a** the GCV function and **b** the RGCV ($\gamma = 0.1$) function for Ex. 1 with $\sigma = 0.1$ and $n = 20$. The corresponding GCV and RGCV estimates are marked with a + symbol

For 1000 replicates of the data, Fig. 3a shows the histogram of inefficiencies $I_T(\hat{\lambda}_V) = T(\hat{\lambda}_V)/\min T(\lambda)$ for the GCV estimates $\hat{\lambda}_V$. The histogram has darker shading for any replicate for which the value of $I_T(\hat{\lambda}_V)$ is within 1 % of the value of

$$I_T(0^+) = \frac{T(0^+)}{\min T(\lambda)} = \frac{n^{-1}\sum_{i=1}^{n}[y_i - f(x_i)]^2}{\min T(\lambda)}$$

for the replicate. Here $I_T(0^+)$ is the inefficiency for the extreme undersmoothed estimate $f_{0^+} = \lim_{\lambda\to 0^+} f_\lambda$, i.e. the natural cubic spline interpolating the noisy data $(x_i, y_i)$, $i = 1, \ldots, n$. There are 133 such replicates, so, in about 13.3 % of the replicates, the GCV estimate is far too small, giving a wiggly spline estimate that comes close to interpolating the data.

For the same 1000 replicates of the data, Fig. 3b shows the corresponding histogram of inefficiencies $I_T$ for the RGCV estimates with $\gamma = 0.1$. Now only 4.2 % of the replicates have darker shading, which is a significant improvement over GCV.

As shown in Lukas (2014), the prediction error can be a misleading performance measure because of its poor discrimination of extreme undersmoothing. The histogram in Fig. 3a illustrates this weakness. Focusing on the histogram near 1.4, there are several replicates in the dark shaded area for which the GCV spline estimate has $I_T \approx 1.4$ and yet these estimates would be regarded as much less accurate by eye (undersmoothed) compared to other estimates with $I_T \approx 1.4$ in the light shaded area. Hence, the prediction error does not adequately detect the poor performance of GCV for the replicates in the dark shaded area, and (comparing Fig. 3a, b) it does not reflect the improved performance of RGCV.

The Sobolev error $W(\lambda)$ is significantly better than the prediction error in discriminating extreme undersmoothing (Lukas 2014). Consequently, it clearly detects the poor undersmoothed GCV estimates. To illustrate this, Table 1 shows, for five replicates of the data, the values of the inefficiencies $I_T$ and $I_W$ for the GCV and RGCV (with $\gamma = 0.1$) choices of $\lambda$. For replicate 2, we would regard the RGCV spline, plotted as the dashed curve in Fig. 1a, as much more accurate than the (very





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

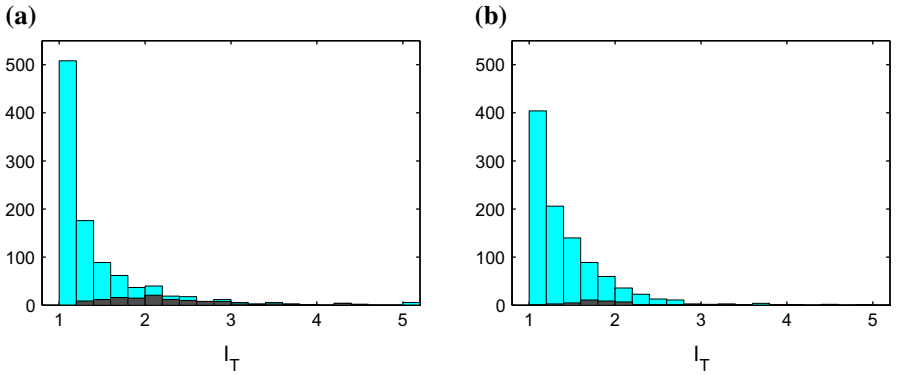


**Fig. 3** Histogram of inefficiencies $I_T$ for **a** GCV and **b** RGCV ($\gamma = 0.1$) for Ex. 1 with $\sigma = 0.1$ and $n = 20$, with *darker shading* where $I_T$ is within 1 % of its extreme undersmoothing value

**Table 1** Values of $I_T$ and $I_W$ for five GCV and RGCV ($\gamma = 0.1$) estimates

| Replicate | $I_T$(GCV) | $I_T$(RGCV) | $I_W$(GCV) | $I_W$(RGCV) |
|---|---|---|---|---|
| 1 | 1.275 | 1.671 | 1.918 | 2.239 |
| 2 | 2.500 | 1.907 | 24.338 | 2.258 |
| 3 | 1.121 | 1.204 | 1.135 | 1.180 |
| 4 | 1.092 | 1.031 | 1.182 | 1.235 |
| 5 | 1.181 | 1.877 | 1.075 | 1.285 |

undersmoothed) GCV spline, plotted as the dotted curve in Fig. 1a. However, this is not reflected in the $I_T$ values in Table 1; $I_T(\text{GCV}) = 2.500$ is only about 31 % larger than $I_T(\text{RGCV}) = 1.907$. The $I_W$ values $I_W(\text{GCV}) = 24.338$ and $I_W(\text{RGCV}) = 2.258$ are a much better reflection of the visual difference in accuracy. In the other direction, for replicate 5 in Table 1, the value $I_T(RGCV) = 1.877$ is about 59 % larger than $I_T(GCV) = 1.181$ (i.e. more than the percentage difference for replicate 2), and yet in Fig. 1b we would not consider the GCV and RGCV splines to be that different. The $I_W$ values are more consistent with this view.

Figure 4a, b show the histograms of $I_W$ for GCV and RGCV (with $\gamma = 0.1$), respectively, for the same 1000 replicates of the data as in Fig. 3a, b. The replicates with $I_W > 10$ are included in the bin at 10. Unlike the histogram in Fig. 3a, the histogram in Fig. 4a clearly identifies the cases where the GCV criterion results in extreme undersmoothing. As mentioned earlier, this occurs for 10–15 % of the replicates. By contrast, from Fig. 4b, RGCV (with $\gamma = 0.1$) performs well for almost all the replicates.

## 4 Optimal value of $\gamma$ for RGCV and a simple rule for choosing $\gamma$

Our aim is to choose the robustness parameter $\gamma$ for RGCV in such a way that it gives reliable and accurate spline estimates. For the reasons above, we will use the loss





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

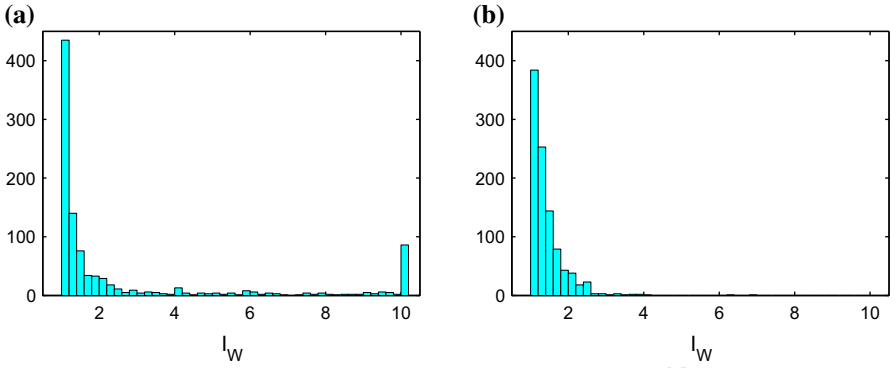


**Fig. 4** Histogram of inefficiencies $I_W$ for **a** GCV and **b** RGCV ($\gamma = 0.1$) for Ex. 1 with $\sigma = 0.1$ and $n = 20$

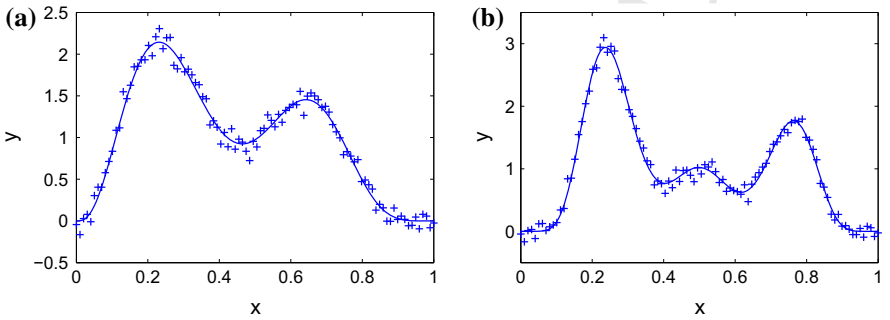


**Fig. 5** Function $f(x)$ (*solid*) and data (+) for **a** Ex. 2 and **b** Ex. 3 with $\sigma = 0.1$ and $n = 100$

function $W(\lambda)$ to define an optimal value of $\gamma$. Theoretically, an appropriate objective is to minimize the expected value of $I_W$ for $\gamma \in [0, 1]$. We will use this objective empirically and define the optimal value of $\gamma$ as the minimizer of the sample mean of the $I_W$ values resulting from 1000 replicates of the data.

First we consider Ex. 1–3 from Craven and Wahba (1979) defined in (9). The graph of the function $f(x)$ (and data) is shown in Figs. 1a and 5a, b, respectively.

For each example, we define 21 cases corresponding to the values $n = 25$, 50, 100, 200, 400, 800, 1600 and $\sigma = 0.05$, 0.1, 0.2. For each case, we generated 1000 replicates of the data, and for each replicate we computed the RGCV spline estimates for $\gamma = 0$, 0.05, 0.1, 0.2, ... ,0.9, 1 and their inefficiencies $I_W$. From the 1000 replicates, we computed the mean of the inefficiencies, denoted mean($I_W$), and the standard error.

For Ex. 1, Fig. 6a, b display the results for $n = 25$ and $n = 1600$, respectively. The figures show mean($I_W$) as a function of $\gamma$ for the three $\sigma$ values 0.05 (dotted), 0.1 (dashed) and 0.2 (solid). Also plotted are error bars for 1 standard error above and below mean($I_W$). For $n$ between 25 and 1600, the plots of mean($I_W$) are similar to and 'between' those in Fig. 6a, b. For all the cases, the value of mean($I_W$) for RGCV with $\gamma \in (0.3, 1)$ (approximately) is less than that for GCV, which corresponds to $\gamma = 1$.



Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

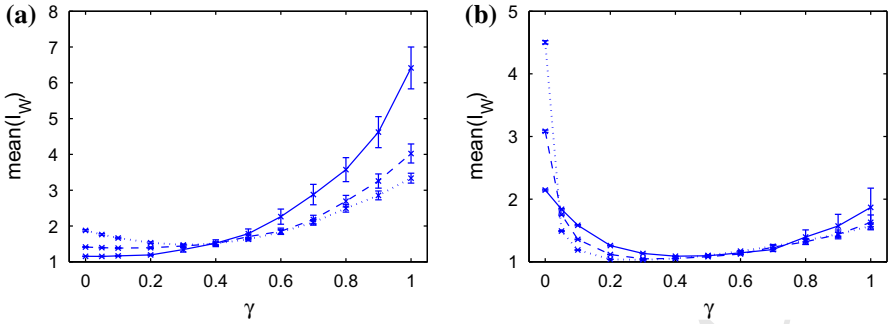


**Fig. 6** Plot of mean($I_W$) versus $\gamma$ for Ex. 1 with **a** $n = 25$ and **b** $n = 1600$, and $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*)

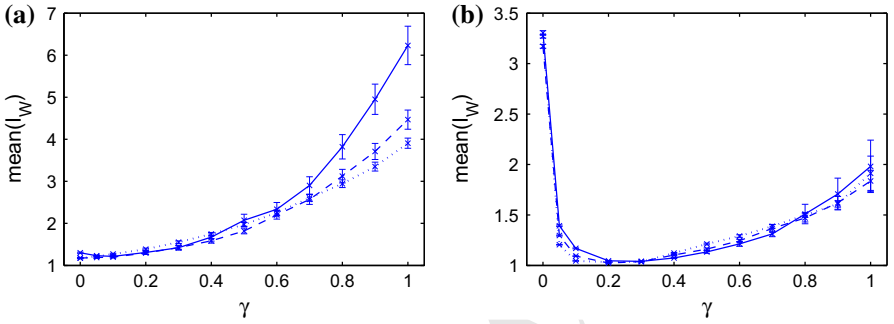


**Fig. 7** Plot of mean($I_W$) versus $\gamma$ for Ex. 2 with **a** $n = 25$ and **b** $n = 1600$, and $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*)

This improvement is most pronounced for smaller $n$, where GCV is less reliable. Note that the graphs in Fig. 6b have a similar shape to the asymptotic curves in Fig. 4a of Lukas et al. (2012).

The same behavior for mean($I_W$) holds for Ex. 2, for which the corresponding figures are Fig. 7a for n = 25 and Fig. 7b for n = 1600.

For Ex. 3, with the corresponding Fig. 8a for n = 25 and Fig. 8b for n = 1600, the same behavior for mean($I_W$) holds except that the values of mean($I_W$) for GCV ($\gamma = 1$) with $n = 1600$ are larger than for Ex. 1 and Ex. 2, and RGCV shows similar improvement with $\gamma$ decreasing in (0.3, 1) for $n = 1600$ as for $n = 25$. This may be explained intuitively by the fact that the function $f(x)$ in Ex. 3 is more rough than those in Ex. 1 and Ex. 2, and so GCV finds it harder to distinguish signal from noise even for large $n$.

For each value of $n$, we determined the optimal value $\gamma_{\text{opt}}$ (that minimizes mean($I_W$) for $\gamma = 0, 0.05, 0.1, 0.2, \ldots, 0.9, 1$) and the minimum value mean($I_W$)($\gamma_{\text{opt}}$) for the three values of $\sigma$. For Ex. 1, Fig. 9a shows plots of $\gamma_{\text{opt}}$ versus $n$ for $\sigma = 0.05$ (dotted), $\sigma = 0.1$ (dashed) and $\sigma = 0.2$ (solid). Figures 10a and 11a show the corresponding plots for Ex. 2 and Ex. 3, respectively.

The graphs of mean($I_W$), as a function of $\gamma$, are often very flat near the minimum. In addition to $\gamma_{\text{opt}}$, we also find for each $n$ the interval of $\gamma$ values for which mean($I_W$)



Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

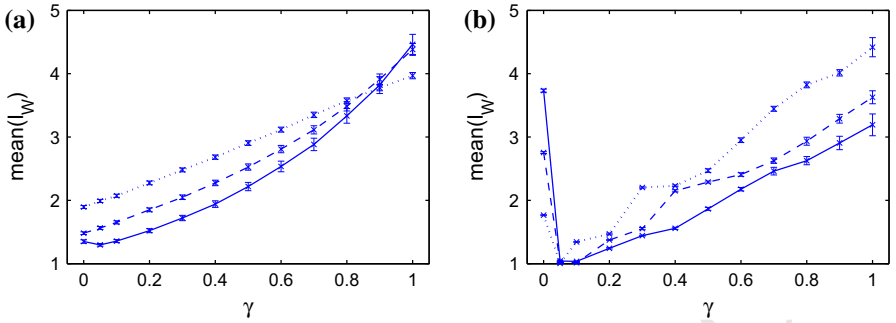


**Fig. 8** Plot of mean($I_W$) versus $\gamma$ for Ex. 3 with **a** $n = 25$ and **b** $n = 1600$, and $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*)

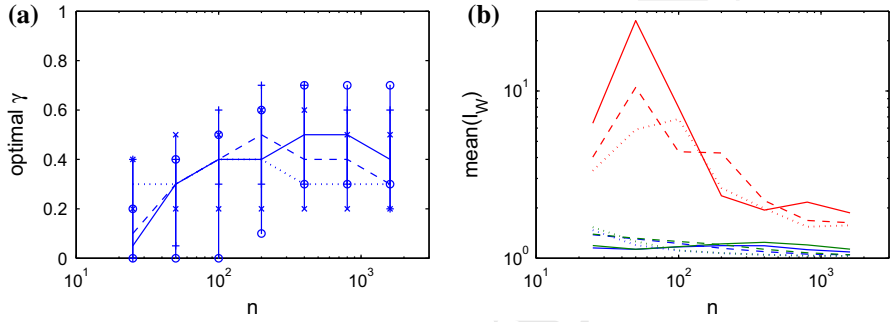


**Fig. 9** Results for Ex. 1 with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*): **a** plot of $\gamma_{\text{opt}}$ versus $n$ and **b** plot of mean($I_W$) versus $n$ for $\gamma = \gamma_{\text{opt}}$ (*lowest*), $\gamma = \gamma_{\text{rule}}$ (*middle*) and $\gamma = 1$ (i.e. GCV) (*highest*)

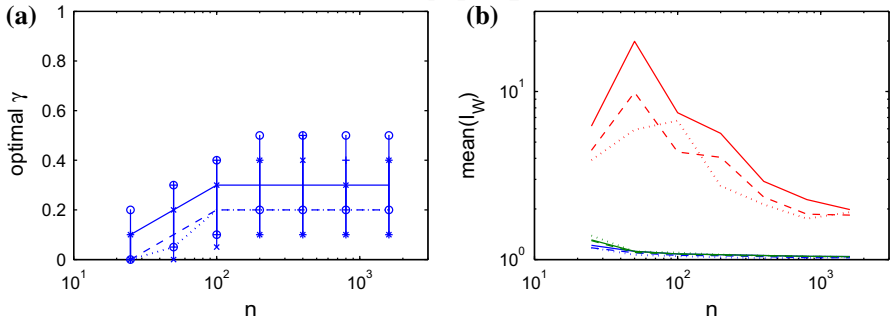


**Fig. 10** Results for Ex. 2 with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*): **a** plot of $\gamma_{\text{opt}}$ versus $n$ and **b** plot of mean($I_W$) versus $n$ for $\gamma = \gamma_{\text{opt}}$ (*lowest*), $\gamma = \gamma_{\text{rule}}$ (*middle*) and $\gamma = 1$ (i.e. GCV) (*highest*)

is within 10 % of its minimum, and these intervals are included in Figs. 9a, 10a and 11a, with endpoints denoted by the symbols x, + and ○ for $\sigma = 0.05$, $\sigma = 0.1$ and $\sigma = 0.2$, respectively.



Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

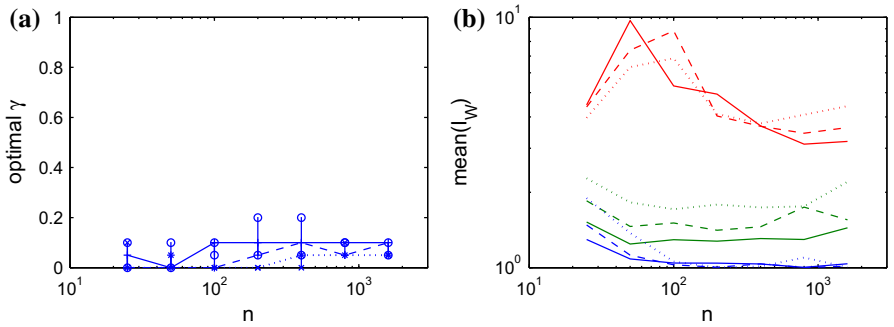


**Fig. 11** Results for Ex. 3 with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*): **a** plot of $\gamma_{\text{opt}}$ versus $n$ and **b** plot of mean($I_W$) versus $n$ for $\gamma = \gamma_{\text{opt}}$ (*lowest*), $\gamma = \gamma_{\text{rule}}$ (*middle*) and $\gamma = 1$ (i.e. GCV) (*highest*)

In Figs. 9a, 10a and 11a, for each value of $\sigma$, there is a clear trend that $\gamma_{\text{opt}}$ increases with $n$ up to about $n = 100$ (or $n = 400$ in Fig. 11a) and then levels off. Moreover, in Figs. 9a and 10a there are large intervals about $\gamma_{\text{opt}}$ in which the value of $\gamma$ is close to optimal.

Based on the behavior of $\gamma_{\text{opt}}$ in Figs. 9a, 10a and 11a, and also on the behavior of the asymptotic inefficiencies $I_T$ and $I_W$ in Figs. 4(a) and 5(a) of Lukas et al. (2012), we propose the following simple rule to choose the value of $\gamma$:

$$\gamma_{\text{rule}} = \begin{cases} 0.2, & \text{if } n < 100, \\ 0.3, & \text{if } n \geq 100. \end{cases} \tag{14}$$

For Ex. 1, 2 and 3, respectively, Figs. 9b, 10b and 11b show mean($I_W$) plotted against $n$ for $\gamma = \gamma_{\text{opt}}$ (lowest curve), $\gamma = \gamma_{\text{rule}}$ (middle curve) and $\gamma = 1$ (i.e. GCV) (highest curve), for the three values $\sigma = 0.05$ (dotted), $\sigma = 0.1$ (dashed) and $\sigma = 0.2$ (solid). Note that, for the most part, the optimal value of mean($I_W$), i.e. for $\gamma = \gamma_{\text{opt}}$, is very close to the ideal value of 1. Furthermore, the use of the rule (14) results in a value of mean($I_W$) that is not far from optimal, especially in Ex. 2. Also, in all three examples the value of mean($I_W$) for RGCV with $\gamma = \gamma_{\text{rule}}$ is much smaller than that resulting from GCV. The difference is greatest for small and medium sized values of $n$ (say $n \leq 200$), but there is still a significant difference for much larger values of $n$.

To test the rule in (14), we consider two more examples—the Motorcycle function (Cummins et al. 2001), defined in (10) and shown in Fig. 12a, and the peaks function, defined in (11) and shown in Fig. 13a. As before, Figs. 12b and 13b show mean($I_W$) plotted against $n$ for $\gamma = \gamma_{\text{opt}}$ (lowest curve), $\gamma = \gamma_{\text{rule}}$ (middle curve) and $\gamma = 1$ (i.e. GCV) (highest curve), for the three values $\sigma = 0.05$ (dotted), $\sigma = 0.1$ (dashed) and $\sigma = 0.2$ (solid). As in Ex. 1–3, RGCV with $\gamma = \gamma_{\text{rule}}$ in (14) performs well again. For all $n$ and $\sigma$, the value of mean($I_W$) for the rule is usually close to the optimal value, and it is usually much less than that for GCV.

Because the rule (14) for choosing $\gamma$ in RGCV is consistently close to giving the optimal value of mean($I_W$) for the examples considered, we can conclude that RGCV





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

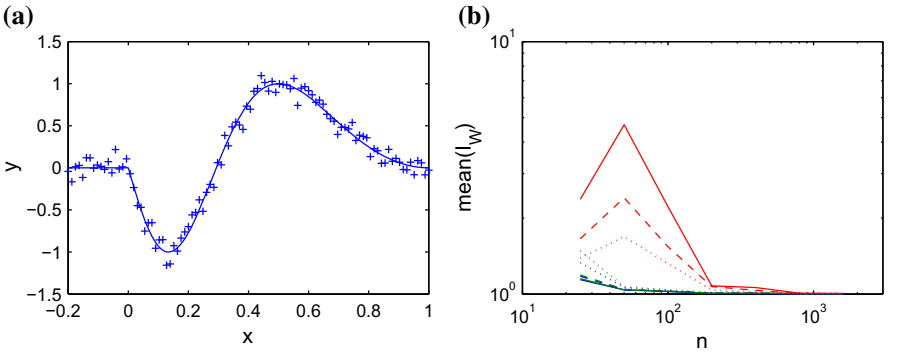


**Fig. 12** **a** Motorcycle function (*solid*) and data (+) for $\sigma = 0.1$ and $n = 100$ **b** Plot of mean($I_W$) versus $n$, with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*), for $\gamma = \gamma_{\text{opt}}$ (*lowest*), $\gamma = \gamma_{\text{rule}}$ (*middle*) and $\gamma = 1$ (i.e. GCV) (*highest*)

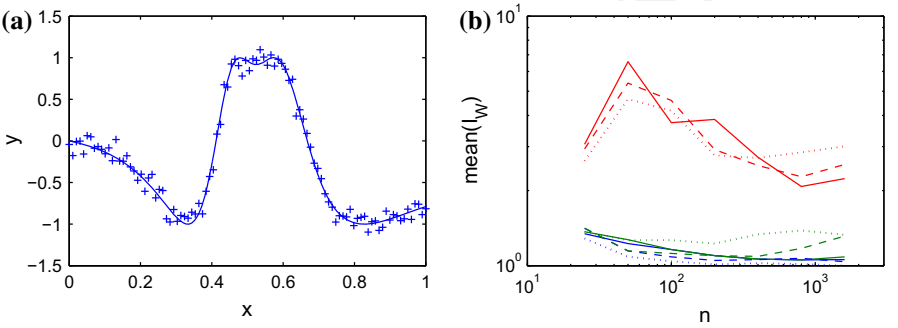


**Fig. 13** **a** Peaks function (*solid*) and data (+) for $\sigma = 0.1$ and $n = 100$ **b** Plot of mean($I_W$) versus $n$, with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*), for $\gamma = \gamma_{\text{opt}}$ (*lowest*), $\gamma = \gamma_{\text{rule}}$ (*middle*) and $\gamma = 1$ (i.e. GCV) (*highest*)

with $\gamma = \gamma_{\text{rule}}$ will perform well over a reasonably wide class of problems with different sample sizes.

## 5 Optimal value of $\rho$ for modified GCV and a simple rule for choosing $\rho$

In this section, we will present simulation results for the modified GCV criterion defined by (5), and consider the optimal choice of the parameter $\rho$. We use the same test examples and objective as for RGCV in Sect. 4, i.e. we minimize mean($I_W$) for the same 1000 replicates of the data.

For Ex. 1, Fig. 14a, b display the results for $n = 25$ and $n = 1600$, respectively. These figures (which correspond to Fig. 6a, b) show mean($I_W$) as a function of $\rho$ for the three $\sigma$ values 0.05 (dotted), 0.1 (dashed) and 0.2 (solid). Also plotted are error bars for 1 standard error above and below mean($I_W$). For $n$ between 25 and 1600, the plots of mean($I_W$) are similar to and 'between' those shown in Fig. 14a, b. For all the cases, the value of mean($I_W$) for modified GCV, with $\rho$ in a certain subinterval of



Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

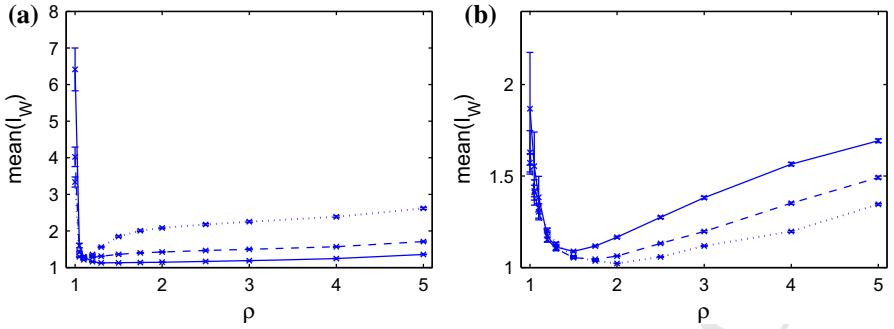


**Fig. 14** Plot of mean($I_W$) versus $\rho$ in modified GCV for Ex. 1 with **a** $n = 25$ and **b** $n = 1600$, and $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*)

(1, 5), is less than that for GCV, which corresponds to $\rho = 1$. Note that the graphs in Fig. 14b have a similar shape to the asymptotic curves in Fig. 5b of Lukas et al. (2012).

For each value of $n$, we find the optimal value $\rho_{\text{opt}}$ (that minimizes mean($I_W$) for $\rho = 1, 1.05, 1.1, 1.2, 1.3, 1.5, 1.75, 2, 2.5, 3, 4, 5$) and the minimum value mean($I_W$)($\rho_{\text{opt}}$) for the three values of $\sigma$. For Ex. 1, Fig. 15a shows plots of $\rho_{\text{opt}}$ versus $n$ for $\sigma = 0.05$ (dotted), $\sigma = 0.1$ (dashed) and $\sigma = 0.2$ (solid). Figures 16a and 17a show the corresponding plots for Ex. 2 and Ex. 3, respectively. In addition to $\rho_{\text{opt}}$, we also find for each $n$ the interval of $\rho$ values for which mean($I_W$) is within 10 % of its minimum, and these intervals are included in Figs. 15a, 16a and 17a, with endpoints denoted by the symbols x, + and o for $\sigma = 0.05$, $\sigma = 0.1$ and $\sigma = 0.2$, respectively.

There is an interesting difference in the behavior of $\rho_{\text{opt}}$ for modified GCV compared to $\gamma_{\text{opt}}$ for RGCV. One might expect that $\rho_{\text{opt}}$ would be larger for smaller $n$ (in the same way that $1/\gamma_{\text{opt}}$ is larger for smaller $n$). However, in Figs. 15a, 16a and 17a, there is a clear trend (especially in the last two figures) that $\rho_{\text{opt}}$ increases as $n$ increases. This difference may be explained intuitively by the fact that modified GCV uses a more strict form of stabilization than RGCV. In fact, for any $\rho > 1$, the smoothing parameter $\lambda$ is constrained to be greater than the value for which the effective degrees of freedom tr$A$ equals $n/\rho$. For the small sample size $n = 25$, the use of the relatively large value $\rho = 5$ reduces the effective degrees of freedom to a maximum of 5. This does not allow enough flexibility for a good fit, especially for the bimodal and trimodal functions in Ex. 2 and 3, respectively. The strict constraint involved in modified GCV is not as serious an issue for large $n$ because, for larger values of $\rho$, there are still sufficiently many effective degrees of freedom.

Based on the behavior of $\rho_{\text{opt}}$ in Figs. 15a, 16a and 17a, and also on the behavior of the asymptotic inefficiencies $I_T$ and $I_W$ in Figs. 4b and 5b of Lukas et al. (2012), we propose the following simple rule to choose the value of $\rho$:

$$\rho_{\text{rule}} = \begin{cases} 1.3, & \text{if } n < 100, \\ 2, & \text{if } n \geq 100. \end{cases} \tag{15}$$

Note that the rules (15) and (14) are consistent for large $n$, since, by (6), the value $\rho = 2$ in (15) corresponds to $\gamma = 3/11 \approx 0.3$.





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

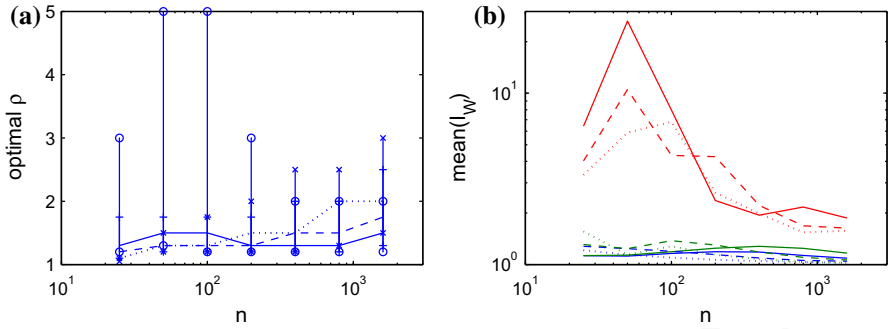


**Fig. 15** Results for modified GCV on Ex. 1 with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*): **a** plot of $\rho_{\text{opt}}$ versus $n$ and **b** plot of mean$(I_W)$ versus $n$ for $\rho = \rho_{\text{opt}}$ (*lowest*), $\rho = \rho_{\text{rule}}$ (*middle*) and $\rho = 1$ (i.e. GCV) (*highest*)

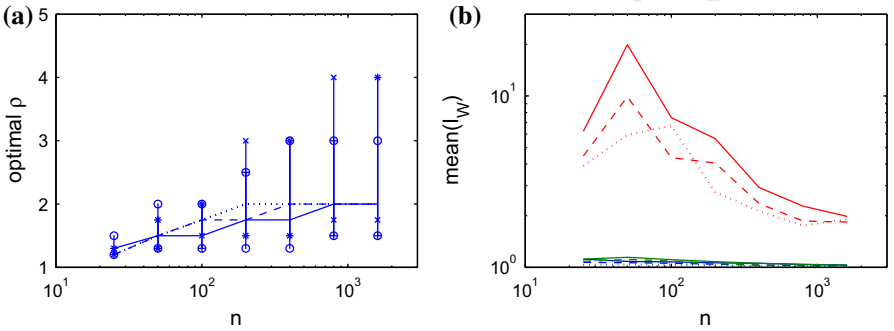


**Fig. 16** Results for modified GCV on Ex. 2 with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*): **a** plot of $\rho_{\text{opt}}$ versus $n$ and **b** plot of mean$(I_W)$ versus $n$ for $\rho = \rho_{\text{opt}}$ (*lowest*), $\rho = \rho_{\text{rule}}$ (*middle*) and $\rho = 1$ (i.e. GCV) (*highest*)

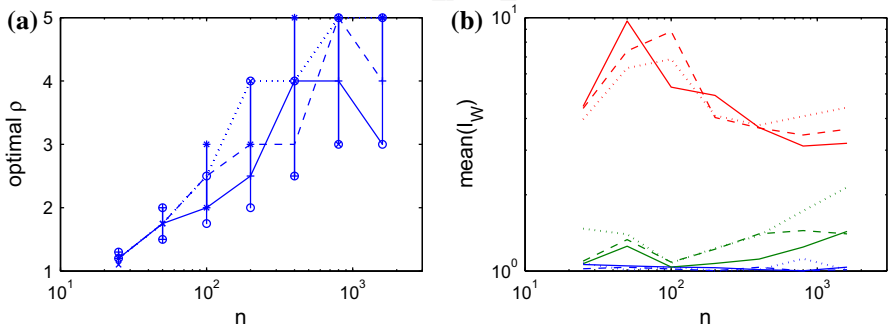


**Fig. 17** Results for modified GCV on Ex. 3 with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*): **a** plot of $\rho_{\text{opt}}$ versus $n$ and **b** plot of mean$(I_W)$ versus $n$ for $\rho = \rho_{\text{opt}}$ (*lowest*), $\rho = \rho_{\text{rule}}$ (*middle*) and $\rho = 1$ (i.e. GCV) (*highest*)

For Ex. 1, 2 and 3, respectively, Figs. 15b, 16b and 17b show mean$(I_W)$ plotted against $n$ for $\rho = \rho_{\text{opt}}$ (lowest curve), $\rho = \rho_{\text{rule}}$ (middle curve) and $\rho = 1$ (i.e. GCV) (highest curve), for the three values $\sigma = 0.05$ (dotted), $\sigma = 0.1$ (dashed) and $\sigma = 0.2$



Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

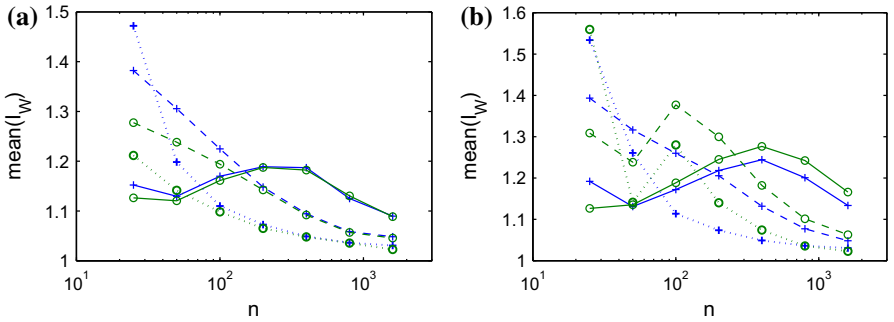


**Fig. 18** Plot of mean($I_W$) versus $n$ for RGCV and modified GCV with **a** $\gamma = \gamma_{\text{opt}}$ (+) and $\rho = \rho_{\text{opt}}$ (o) and with **b** $\gamma = \gamma_{\text{rule}}$ (+) and $\rho = \rho_{\text{rule}}$ (o) for Ex. 1 with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*)

(solid). From these figures, the same conclusions hold as for the corresponding plots for RGCV in Figs. 9b, 10b and 11b. For the most part, the optimal value of mean($I_W$), i.e. for $\rho = \rho_{\text{opt}}$, is very close to the ideal value of 1. Furthermore, the use of the rule (15) results in a value of mean($I_W$) that is not far from optimal, especially in Ex. 2. Also, in all three examples, the value of mean($I_W$) for modified GCV with $\rho = \rho_{\text{rule}}$ is much smaller than that resulting from GCV. The difference is greatest for small and medium sized values of $n$ (say $n \leq 200$), but there is still a significant difference for much larger values of $n$.

The rule in (15) was also tested on the Motorcycle function (Cummins et al. 2001) defined in (10) and the peaks function defined in (11). As in Ex. 1–3, modified GCV with $\rho = \rho_{\text{rule}}$ in (15) performs well for these examples (the plots are omitted). For all $n$ and $\sigma$, the value of mean($I_W$) for the rule is usually close to the optimal value, and it is usually much less than that for GCV.

Because the rule (15) for choosing $\rho$ in modified GCV is consistently close to yielding the optimal value of mean($I_W$) for all the examples considered, we can conclude that modified GCV with $\rho = \rho_{\text{rule}}$ will perform well over a reasonably wide class of problems with different sample sizes.

## 6 Comparison of RGCV and modified GCV

In this last section, we will compare RGCV and modified GCV for the five test examples. Figures 18, 19, 20, 21 and 22 correspond to Ex. 1–3, the Motorcycle function and the peaks function, respectively. For each of these figures, panel (a) shows the value of mean($I_W$) for RGCV with $\gamma = \gamma_{\text{opt}}$ (+ symbol) and modified GCV with $\rho = \rho_{\text{opt}}$ (○ symbol) plotted against $n$ for the three values $\sigma = 0.05$ (dotted), $\sigma = 0.1$ (dashed) and $\sigma = 0.2$ (solid). This means that each graph represents the optimal value of mean($I_W$) that can be achieved by RGCV or by modified GCV.

Allowing for the coarseness of the grids used for the parameters $\gamma$ and $\rho$, the graphs of the optimal value of mean($I_W$) for RGCV and modified GCV are quite close for $n \geq 200$. This is consistent with the asymptotic result that RGCV and modified GCV





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

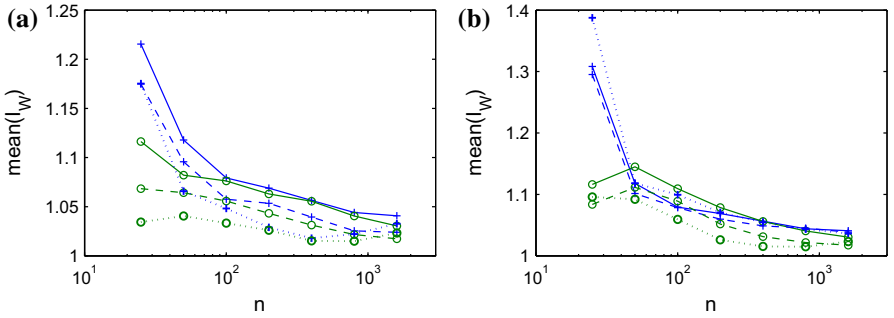


**Fig. 19** Plot of mean($I_W$) versus $n$ for RGCV and modified GCV with **a** $\gamma = \gamma_{\text{opt}}$ (+) and $\rho = \rho_{\text{opt}}$ (o) and with **b** $\gamma = \gamma_{\text{rule}}$ (+) and $\rho = \rho_{\text{rule}}$ (o) for Ex. 2 with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*)

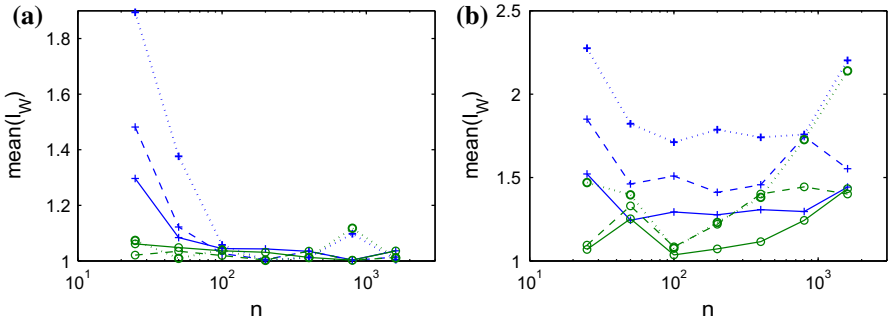


**Fig. 20** Plot of mean($I_W$) versus $n$ for RGCV and modified GCV with **a** $\gamma = \gamma_{\text{opt}}$ (+) and $\rho = \rho_{\text{opt}}$ (o) and with **b** $\gamma = \gamma_{\text{rule}}$ (+) and $\rho = \rho_{\text{rule}}$ (o) for Ex. 3 with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*)

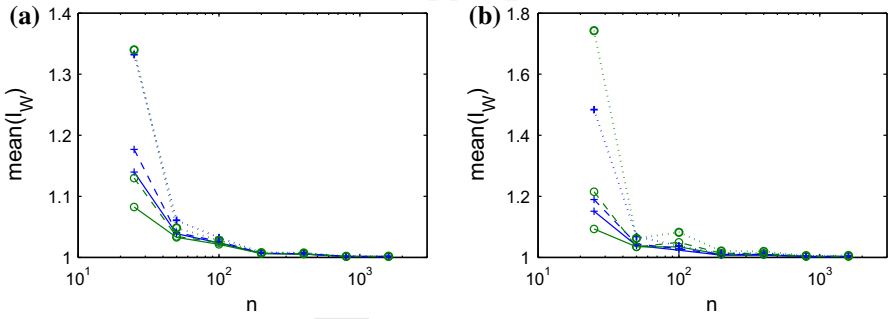


**Fig. 21** Plot of mean($I_W$) versus $n$ for RGCV and modified GCV with **a** $\gamma = \gamma_{\text{opt}}$ (+) and $\rho = \rho_{\text{opt}}$ (o) and with **b** $\gamma = \gamma_{\text{rule}}$ (+) and $\rho = \rho_{\text{rule}}$ (o) for the Motorcycle function with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*)

have the same behavior as $n \to \infty$ (Cummins et al. 2001; Lukas 2008). However, there are differences between the graphs for RGCV and modified GCV for smaller values of $n$. In almost all cases, modified GCV gives a smaller optimal value of mean($I_W$)



Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

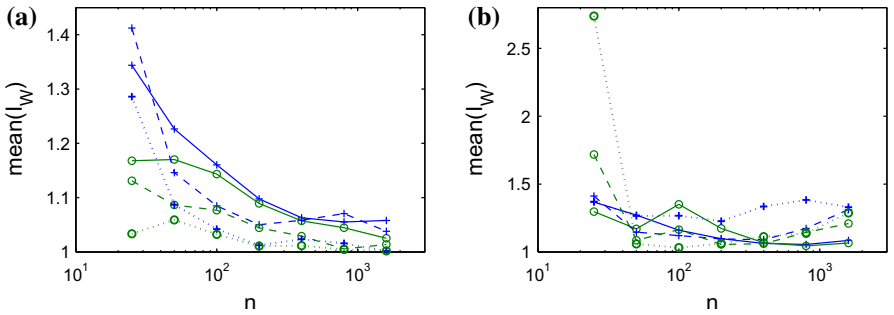


**Fig. 22** Plot of mean($I_W$) versus $n$ for RGCV and modified GCV with **a** $\gamma = \gamma_{\text{opt}}$ (+) and $\rho = \rho_{\text{opt}}$ (o) and with **b** $\gamma = \gamma_{\text{rule}}$ (+) and $\rho = \rho_{\text{rule}}$ (o) for the peaks function with $\sigma = 0.05$ (*dotted*), $\sigma = 0.1$ (*dashed*) and $\sigma = 0.2$ (*solid*)

for small $n$ than does RGCV. While this is an important observation, it should also be kept in mind that the optimal values of the parameters $\gamma$ and $\rho$ will not be known in practice.

Figures 18b, 19, 20, 21 and 22b compare RGCV and modified GCV with the two practical rules (14) and (15) for choosing their parameters $\gamma$ and $\rho$, respectively. In these figures, the general performance of RGCV and modified GCV is about the same. There is quite a bit of variation in the graphs; in some situations, RGCV performs better, and, in other situations, modified GCV performs better. However, it can be seen that, in almost all situations, both RGCV and modified GCV with the rules (14) and (15) yield values of mean($I_W$) that are less than 1.5, and, in most situations, the values are less than 1.3. Therefore, both methods usually perform very well with respect to the Sobolev error, and considerably better than GCV, as seen in the graphs for GCV in Sects. 4 and 5.

## References


Anselone PM, Laurent PJ (1968) A general method for the construction of interpolating or smoothing spline-functions. Numer Math 12:66–82

Cox DD (1984) Gaussian approximation of smoothing splines. Tech. rep., Dept. Statist., University of Wisconsin/Madison

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31:377–403

Cummins DJ, Filloon TG, Nychka D (2001) Confidence intervals for nonparametric curve estimates: toward more uniform pointwise coverage. J Am Stat Assoc 96:233–246

de Hoog FR, Anderssen RS, Lukas MA (2011) Differentiation of matrix functionals using triangular factorization. Math Comput 80(275):1585–1600

Efron B (2001) Selection criteria for scatterplot smoothers. Ann Stat 29:470–504

Eubank RL (1988) Spline smoothing and nonparametric regression. Dekker, New York

Girard DA (2010) Estimating the accuracy of (local) cross-validation via randomised GCV choices in kernel or smoothing spline regression. J Nonparametric Stat 22:41–64

Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman & Hall, London

Gu C (2002) Smoothing spline ANOVA models. Springer, New York

Hutchinson MF, de Hoog FR (1985) Smoothing noisy data with spline functions. Numer Math 47:99–106

Kim YJ, Gu C (2004) Smoothing spline Gaussian regression: more scalable computation via efficient approximation. J R Stat Soc Ser B 66:337–356
Li KC (1986) Asymptotic optimality of $C_L$ and generalized cross-validation in ridge regression with application to spline smoothing. Ann Stat 14:1101–1112
Lukas MA (2006) Robust generalized cross-validation for choosing the regularization parameter. Inverse Probl 22:1883–1902
Lukas MA (2008) Strong robust generalized cross-validation for choosing the regularization parameter. Inverse Probl 24(034):006
Lukas MA (2014) Performance criteria and discrimination of extreme undersmoothing in nonparametric regression. J Stat Plan Inference 153:56–74
Lukas MA, de Hoog FR, Anderssen RS (2010) Efficient algorithms for robust generalized cross-validation spline smoothing. J Comput Appl Math 235:102–107
Lukas MA, de Hoog FR, Anderssen RS (2012) Performance of robust GCV and modified GCV for spline smoothing. Scand J Stat 39:97–115
Nychka D (1990) The average posterior variance of a smoothing spline and a consistent estimate of the average squared error. Ann Stat 18:415–428
Reinsch CH (1967) Smoothing by spline functions. Numer Math 10:177–183
Reinsch CH (1971) Smoothing by spline functions II. Numer Math 16:451–454
Robinson T, Moyeed R (1989) Making robust the cross-validatory choice of smoothing parameter in spline smoothing regression. Commun Stat Theory Methods 18:523–539
Wahba G (1990) Spline models for observational data. SIAM, Philadelphia





Journal: 180 Article No.: 0577 TYPESET DISK LE CP Disp.:2015/4/9 Pages: 21 Layout: Small-X

Journal: 180
Article: 577

Springer
the language of science

Author Proof

# Author Query Form

**Please ensure you fill out your response to the queries raised below and return this form along with your corrections**

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

| Query | Details required | Author's response |
|---|---|---|
| 1. | Please check and confirm if the authors and their respective affiliations have been correctly identified. Amend if necessary. | |