Performance of Robust GCV and Modified GCV for Spline Smoothing

MARK A. LUKAS
Mathematics and Statistics, Murdoch University
FRANK R. DE HOOG and ROBERT S. ANDERSSEN
CSIRO Mathematics, Informatics and Statistics

ABSTRACT. While it is a popular selection criterion for spline smoothing, generalized cross-validation (GCV) occasionally yields severely undersmoothed estimates. Two extensions of GCV called robust GCV (RGCV) and modified GCV have been proposed as more stable criteria. Each involves a parameter that must be chosen, but the only guidance has come from simulation results. We investigate the performance of the criteria analytically. In most studies, the mean square prediction error is the only loss function considered. Here, we use both the prediction error and a stronger Sobolev norm error, which provides a better measure of the quality of the estimate. A geometric approach is used to analyse the superior small-sample stability of RGCV compared to GCV. In addition, by deriving the asymptotic inefficiency for both the prediction error and the Sobolev error, we find intervals for the parameters of RGCV and modified GCV for which the criteria have optimal performance.

Key words: asymptotic, generalized cross-validation, nonparametric regression, small sample, Sobolev error, spline smoothing

Running headline: Robust GCV and modified GCV criteria

1 Introduction

In many data analysis applications, it is required to fit a smooth curve to noisy data

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

where $f(x)$ is smooth and the random errors ε_i are assumed to be independent and identically distributed with mean 0 and common variance σ^2. Besides yielding good estimates of $f(x)$ at the design points x_i, the curve and its derivative, respectively, should closely track $f(x)$ and $f'(x)$ over the whole interval. Smoothing splines are widely used for this and related nonparametric regression problems; see e.g. Eubank (1988); Gu (2002); Ramsay & Silverman (2005); Wahba (1990). The natural polynomial smoothing spline of degree $2m-1$ is defined as the minimizer f_λ of

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

over all functions f for which $f^{(m)}$ is square integrable. The smoothing parameter $\lambda > 0$ determines the amount of smoothing, and its selection is critical to generating a good spline estimate f_λ.

One of the most popular parameter selection criteria is generalized cross-validation (GCV) (Craven & Wahba, 1979; Wahba, 1990). Denote $\boldsymbol{f}_\lambda = (f_\lambda(x_1), \ldots, f_\lambda(x_n))^T$ and let $A(\lambda)$ be the smoothing matrix defined by $\boldsymbol{f}_\lambda = A(\lambda)\boldsymbol{y}$. The GCV criterion selects λ 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 $\|\cdot\|$ denotes the Euclidean norm. For spline smoothing and also for more general estimation problems, it is known that GCV has good asymptotic properties as the sample size $n \to \infty$. In particular, under mild conditions, the GCV estimate is asymptotically optimal with respect to the mean square prediction error (Li, 1986). However, for small or moderately sized samples, it has been observed (Wahba, 1990, sect. 4.9) that GCV has significant variability, sometimes giving a parameter value that is far too small (possibly even 0), resulting in a very rough spline estimate. Efron (2001) gave a novel geometric interpretation of the erratic small-sample behaviour of GCV and the closely related Mallows C_p criterion. He showed that C_p suffers from an instability (called the reversal effect) in which, for a small change in the data, a desired 'optimal' value of λ can change from being a minimizer of the C_p function to becoming a local maximizer.

Generalized maximum likelihood (GML) is another well-known selection criterion (Wahba, 1985), which is equivalent to restricted maximum likelihood (REML) in the mixed model formulation of smoothing splines (Ruppert *et al.*, 2003). It is shown in Efron (2001); Kou & Efron (2002) that, although GML is more stable than GCV, it can have a large bias. Motivated by the deficiencies of GCV and GML, Kou & Efron (2002) proposed a new criterion, called the extended exponential criterion, which has much less variability than GCV. Recently, Hall & Robinson (2009) showed that the variability of cross-validation (CV) can be significantly reduced by bagging either the CV function or the CV bandwidth estimate.

We consider another selection criterion called robust GCV (RGCV). This criterion was first proposed for spline smoothing by Robinson & Moyeed (1989), who found in simulations that it has much less variability than GCV (see also van der Linde (2000)). In the more general framework of Tikhonov regularization of linear inverse problems (which includes spline smoothing), the RGCV criterion was developed in Lukas (2006, 2008). The criterion uses an approximate average influence measure $F(\lambda) = \mu_2(\lambda)V(\lambda)$, where $\mu_2(\lambda) = n^{-1}\mathrm{tr}(A^2(\lambda))$, and selects λ as the minimizer of the weighted sum

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

where $\gamma \in (0, 1)$ is a robustness parameter. Graphically, the term $(1-\gamma)F(\lambda)$ in $\overline{V}(\lambda)$ changes the shape of the GCV function near 0 so that RGCV is less likely to choose a very small value of λ (Lukas, 2006). For spline smoothing, the function $\overline{V}(\lambda)$ (like $V(\lambda)$) can be computed efficiently in $O(m^2 n)$ operations (Lukas *et al.*, 2010).

In this paper, it is shown that RGCV is a very effective selection criterion for spline smoothing problems of any sample size. We determine the precise effect of the parameter γ and hence find a range of values for which the RGCV estimate is both stable and has good performance. For the loss function, we use both the (mean square) prediction error $T(\lambda) = n^{-1}\|\boldsymbol{f}_\lambda - \boldsymbol{f}\|^2$, where $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))^T$, and a stronger Sobolev error.

In most theoretical or empirical studies of selection criteria, the prediction error is the only loss function considered; see e.g. Kim & Gu (2004); Li (1986). The asymptotic behaviour and optimal rate of the prediction risk $ET(\lambda)$ and its minimizer λ_{ET} are well known (Wahba, 1990). However, the prediction error has limitations, since it is a pointwise measure and, furthermore, it is insensitive to the derivative and curvature of $f_\lambda - f$, which are important for the quality of the fit. This issue is especially important in applications where there is a specific requirement for estimates of $f'(x)$ or $f''(x)$ (Ramsay & Silverman, 2005).

Since f is assumed to be smooth, it is reasonable to use derivatives not only in the roughness penalty in (2) to construct the family of spline estimates $\{f_\lambda\}$, but also in a loss function that defines an optimal estimate from the family, i.e. an optimal value of λ. Suppose that f belongs to the Sobolev space $\mathcal{W} = \mathcal{W}^{m,2}[a,b]$ of functions whose mth weak derivative is in $L_2(a,b)$. Then a natural measure of the accuracy of f_λ is obtained by using the squared Sobolev norm on $\mathcal{W}$ to obtain

$$W(\lambda) := \|f_\lambda - f\|^2_{\mathcal{W}} = \int_a^b (f_\lambda(x) - f(x))^2 dG + \int_a^b (f_\lambda^{(m)}(x) - f^{(m)}(x))^2 dx, \tag{5}$$

which will be called the Sobolev error. Here G is the limit as $n \to \infty$ of the empirical distribution functions for the points $\{x_i\}$. Note that, from (2), f_λ belongs to $\mathcal{W}$ by definition. In fact, f_λ has higher smoothness if $m \geq 2$, since $f_\lambda \in C^{2m-2}[a,b]$ (Wahba, 1990). The asymptotic behaviour and optimal rate of the Sobolev risk $EW(\lambda)$ and its minimizer λ_{EW} are derived in Cox (1984b). We will also consider, as an extension of (5), the weighted Sobolev error $W_\kappa(\lambda) := \|f_\lambda - f\|^2_{L_2(G)} + \kappa\|f_\lambda^{(m)} - f^{(m)}\|^2_{L_2}$, where $\kappa > 0$ is a constant.

In section 2, we investigate the small-sample behaviour of RGCV using the geometric approach of Efron (2001). The reversal effect is used to derive an analytic measure of the instability of RGCV as a function of γ. This function is computed for several examples from the literature. It shows that the stability of RGCV improves considerably as γ is decreased from 1 (the GCV case), and when $\gamma = 0.3$, the criterion is very stable. Simulation results confirm this behaviour.

If γ is decreased too far, the RGCV estimates become too biased, leading to poor spline estimates. In sections 3 and 4, this behaviour is quantified for the large-sample case by deriving the asymptotic inefficiency of the (restricted) RGCV estimate with respect to both the prediction risk $ET(\lambda)$ and the Sobolev risk $EW(\lambda)$. For $ET(\lambda)$, the inefficiency grows monotonically but slowly as γ is decreased from 1, and it is still relatively small (approximately 1.1) when $\gamma = 0.3$. For $EW(\lambda)$, as γ is decreased from 1, (an estimate of) the inefficiency decreases to a unique minimum point, and, when $m = 2$, the corresponding optimal value of γ lies in the interval $(0, 0.6)$. The same result holds for the weighted Sobolev risk $EW_\kappa(\lambda)$, independent of κ. The

special case of the inefficiency for GCV ($\gamma = 1$) is of interest in its own right.

The small-sample and large-sample results in sections 2 – 4 indicate that RGCV with $\gamma \in [0.2, 0.4]$ (approximately) will give both stable and accurate cubic spline estimates for a wide class of functions f. Robinson & Moyeed (1989) found that the value $\gamma = 0.5$ gave good results in simulations, but they did not provide detailed comparisons. A large simulation study in Lukas *et al.* (2008) confirms that RGCV performs well for $\gamma \in [0.2, 0.4]$.

Perhaps the simplest approach to stabilizing GCV is the modified GCV criterion (Cummins *et al.*, 2001; Kim & Gu, 2004). In this criterion, the GCV function $V(\lambda)$ in (3) is modified to the score function

$$V_\rho(\lambda) = \frac{n^{-1}\|(I - A(\lambda))\boldsymbol{y}\|^2}{[n^{-1}\mathrm{tr}(I - \rho A(\lambda))]^2} \tag{6}$$

by replacing $\mathrm{tr}(I - A(\lambda))$ with $\mathrm{tr}(I - \rho A(\lambda))$ for some constant $\rho > 1$. This constrains the effective degrees of freedom $\mathrm{tr}A(\lambda)$ to be less than n/ρ. Simulation results in Cummins *et al.* (2001); Kim & Gu (2004) suggest that, for prediction error loss, $\rho = 1.4$ is a good choice.

It is known (Cummins *et al.*, 2001; Lukas, 2008) that under mild assumptions, if $\mathrm{tr}A \to 0$ and $\mathrm{tr}A/\mathrm{tr}(A^2)$ has a limit M as $n \to \infty$, then RGCV has the same asymptotic behaviour as modified GCV, provided the parameters γ and ρ in the criteria are related by $\gamma^{-1} = 1 + 2(\rho - 1)M$. For cubic spline smoothing, the limit condition holds with $M = 4/3$. Therefore, for large n, the good interval $0.2 \le \gamma \le 0.4$ for RGCV defines the corresponding good interval $2.5 \ge \rho \ge 1.5625$ for the modified GCV criterion. The value $\rho = 1.4$ corresponds to $\gamma = 0.484$.

In the last decade, there has been increased interest in penalized splines and P-splines (Ruppert *et al.*, 2003), which are defined using a spline basis and knot sequence that is much smaller than n. Clearly, from (4), RGCV can be used to select the smoothing parameter for these splines. Because they are spline-like smoothers, the small-sample results for RGCV in section 2 apply to them. In addition, if the number of knots increases sufficiently quickly with n, these splines behave asymptotically like smoothing splines (Claeskens *et al.*, 2009), and we expect that asymptotic results similar to those in sections 3 and 4 will also hold.

2 Geometry and small-sample stability of RGCV

GCV is closely related (Efron, 2001) to the C_p criterion of Mallows (1973), and they behave essentially the same in practice (Craven & Wahba, 1979). The C_p criterion selects λ as the minimizer of

$$C(\lambda) = n^{-1}\|\boldsymbol{y} - A(\lambda)\boldsymbol{y}\|^2 + 2\sigma^2 n^{-1}\mathrm{tr}A(\lambda) - \sigma^2, \tag{7}$$

which is an unbiased estimate of the prediction risk $ET(\lambda) = n^{-1}E\|\boldsymbol{f}_\lambda - \boldsymbol{f}\|^2$. By simple differentiation, the C_p estimate satisfies $r'(\lambda) = -2\sigma^2\mu_1'(\lambda)$, where $r(\lambda) = n^{-1}\|(I - A(\lambda))\boldsymbol{y}\|^2$ and $\mu_1(\lambda) = n^{-1}\mathrm{tr}A(\lambda)$, while the GCV estimate satisfies $r'(\lambda) = -2\hat{\sigma}^2(\lambda)\mu_1'(\lambda)$, where $\hat{\sigma}^2(\lambda) = r(\lambda)/(1 - \mu_1(\lambda))$. Therefore, the GCV estimate satisfies the same equation as the C_p estimate, but with the variance estimate $\hat{\sigma}^2(\lambda)$ in place of σ^2.

Similarly, we show that RGCV is closely related to a robustified version of C_p that we call robust C_p (RC_p), which selects λ as the minimizer of

$$\overline{C}(\lambda) = \gamma C(\lambda) + (1-\gamma)\sigma^2\mu_2(\lambda). \tag{8}$$

The RC_p estimate satisfies $r'(\lambda) = -2\sigma^2\mu_1'(\lambda) - k_\gamma\sigma^2\mu_2'(\lambda)$, where $k_\gamma = (1-\gamma)/\gamma$, while the RGCV estimate satisfies the same equation, but with the variance estimates $\hat{\sigma}^2(\lambda) = r(\lambda)/(1-\mu_1(\lambda))$ and $\tilde{\sigma}^2(\lambda) = r(\lambda)/(1+k_\gamma\mu_2(\lambda))$ in place of σ^2 in the first and second terms, respectively.

The instability of C_p and GCV for small n was explained by Efron (2001) using a simple geometric interpretation. It is well known that the smoothing matrix $A(\lambda)$ has a diagonalization $A(\lambda) = U\mathrm{diag}(a_{\lambda i})U^T$, where U is orthogonal and independent of λ, and $a_{\lambda i} = 1/(1+\lambda\tau_i)$, $i = 1, \ldots, n$, for a certain nondecreasing sequence $\{\tau_i\}$, with $\tau_i = 0$ for $i = 1, \ldots, m$. In fact, the analysis in Efron (2001) and here applies to any spline-like smoother, i.e. a linear smoother in which the smoothing matrix can be diagonalized as $A(\lambda) = U\mathrm{diag}(a_{\lambda i})U^T$, where the diagonal elements $a_{\lambda i}$ have the form $a_{\lambda i} = 1/(1+\lambda\tau_i)$.

The function $C(\lambda)$ in (7) can be simplified by defining $\boldsymbol{z} = U^T\boldsymbol{y}/\sigma$ and $\boldsymbol{g} = U^T\boldsymbol{f}/\sigma$, where z_i has mean g_i and variance 1, and substituting for $\boldsymbol{y}$. Then the C_p estimate $\hat{\lambda}_C$ is the minimizer of

$$l_\lambda(\boldsymbol{u}) = \sum(b_{\lambda i}^2 u_i - 2b_{\lambda i}), \tag{9}$$

where $b_{\lambda i} = 1 - a_{\lambda i}$ and $\boldsymbol{u} = (z_1^2, z_2^2, \ldots, z_n^2)^T$. The sum in (9) and throughout this section is over i with $\tau_i \neq 0$. By simple differentiation, $\hat{\lambda}_C$ satisfies the equation $\dot{\boldsymbol{\eta}}_\lambda^T(\boldsymbol{u} - \boldsymbol{\mu}_\lambda) = 0$, where $\eta_{\lambda i} = -b_{\lambda i}^2$ and $\mu_{\lambda i} = 1/b_{\lambda i} = 1 + (1/\lambda)(1/\tau_i)$. The mapping $\lambda \to \boldsymbol{\mu}_\lambda$ defines a line in $\mathbb{R}^{n-m}$ (called the *line of expectations* by Efron (2001)), where $\boldsymbol{\mu}_\lambda \to \mathbf{1} = (1, 1, \ldots, 1)^T$ as $\lambda \to \infty$ and $\boldsymbol{\mu}_\lambda \to \infty$ as $\lambda \to 0$. From the equation for $\hat{\lambda}_C$, it can be shown (Efron, 2001; Kou & Efron, 2002) that, for a small change in the data, the C_p function can go from having a (local) minimum at a desired 'optimal' value λ_0 to a (local) maximum at λ_0. Then, for such perturbed data, the C_p choice will be far from λ_0. This phenomenon, called the *reversal effect* in Efron (2001), can occur when $\boldsymbol{u}$ is in a certain half-space called the *reversal region*. Let $V_{\lambda_0} = \mathrm{diag}(b_{\lambda_0 i}^{-3}/2)$ and $\beta_{\lambda_0} = \ddot{\eta}_{\lambda_0} V_{\lambda_0}^2 \dot{\eta}_{\lambda_0} / (\dot{\eta}_{\lambda_0} V_{\lambda_0}^2 \dot{\eta}_{\lambda_0})$. Then the reversal region is defined as

$$RR = \{\boldsymbol{w} : R_0(\boldsymbol{w}) < 0\}, \quad R_0(\boldsymbol{w}) = \lambda_0^2[\ddot{l}_{\lambda_0}(\boldsymbol{w}) - \beta_{\lambda_0}\dot{l}_{\lambda_0}(\boldsymbol{w})]. \tag{10}$$

Clearly, if $\boldsymbol{u} \in RR$, then C_p cannot select λ_0 as the estimate (whether $\dot{l}_{\lambda_0}(\boldsymbol{u})$ is 0 or not).

Note. For $R_0(\boldsymbol{w})$, there is an error in the defining equation (39) of Kou & Efron (2002) in that, for the purpose of scaling, one should multiply by λ_0^2 as in (10) rather than divide by λ_0^2. For the same reason, the expressions for the mean $M(R_0)$ and variance $V(R_0)$ in Appendix B of Kou & Efron (2002) should not include the factors λ_0^{-4} and λ_0^{-8}, respectively. (Clearly, the expression in Kou & Efron (2002) for $M(R_0)$ is poorly scaled since $\Sigma a_{\lambda_0 i}^2 b_{\lambda_0 i}$ behaves like $c(\lambda_0)^{-1/(2m)}$ as $n \to \infty$ for some constant c (see theorems 2.3 and 5.1 in Kou (2003) for $m = 2$, where the smoothing parameter equals $n\lambda$ in our notation).)

Because $C(\lambda)$ is an unbiased estimate of the prediction risk $ET(\lambda)$, it is natural to use $\lambda_0 = \lambda_{ET}$ in the definition (10), where λ_{ET} is the minimizer of $ET(\lambda)$. Simulation results in Kou & Efron (2002) for two examples with cubic smoothing splines indicate that R_0 with $\lambda_0 = \lambda_{ET}$ is a good predictor of the instability of C_p. The first example has 61 equally spaced points x_i on $[-1, 1]$ (they are not defined explicitly in Kou & Efron (2002), but we take $x_i = -1 + (i-1)/30$, $i = 1, \ldots, 61$), $f(x) = \sin(\pi(x+1))/(x/2+1)$ and the errors $\varepsilon_i \sim N(0,1)$ (i.e. $\sigma = 1$). For this example with 1000 replicates of the data, Fig. 7(b) in Kou & Efron (2002) shows that the degrees of freedom $df = \mathrm{tr}A(\lambda) = \sum a_{\lambda i}$ for the C_p estimate $\hat{\lambda}_C$ has much greater variability when $R_0(\boldsymbol{u}) < 0$ than when $R_0(\boldsymbol{u}) > 0$.

The same is true for GCV when applied to these data, as shown in Figure 1(a). When $R_0 > 0$, most *df* values for the GCV estimates are near $df = 5.26$, the value of *df* corresponding to $\lambda_0 = \lambda_{ET}$. (In Kou & Efron (2002) this value is $df = 5.18$, and the discrepancy could be due to the choice of x_i or to the search grid used for λ.) Of the 1000 replicates, 19.7% have $R_0 < 0$ and for these the *df* values vary considerably between about 5 and 60.

The RC_p function $\overline{C}(\lambda)$ defined in (8) can be expressed as

$$\begin{aligned}\overline{C}(\lambda) &= \gamma[\sigma^2 n^{-1}\sum(b_{\lambda i}^2 z_i^2 - 2b_{\lambda i}) + \sigma^2] + (1-\gamma)\sigma^2 n^{-1}\sum(1-b_{\lambda i})^2 \\ &= \sigma^2 n^{-1}\{\sum[b_{\lambda i}^2(\gamma z_i^2 + (1-\gamma)) - 2b_{\lambda i}] + (1-\gamma)(n-m)\},\end{aligned}$$

so the RC_p estimate is the minimizer of

$$\bar{l}_\lambda(\boldsymbol{u}) = \sum[b_{\lambda i}^2(\gamma u_i + (1-\gamma)) - 2b_{\lambda i}], \tag{11}$$

where $u_i = z_i^2$ as before. Comparing (11) and (9), it is clear that $\bar{l}_\lambda(\boldsymbol{u}) = l_\lambda(\gamma\boldsymbol{u} + (1-\gamma)\mathbf{1})$ for all λ and $\boldsymbol{u}$. This formula leads to a simple geometric explanation of the stability of the RC_p criterion. First we have the following result.

Theorem 1 *Consider a spline-like smoother with $a_{\lambda i} = 1/(1+\lambda\tau_i)$, where the sequence $\{\tau_i\} \geq 0$ is non-constant and nondecreasing; in particular, a spline smoother. The point* $\mathbf{1}$ *(corresponding to $\lambda \to \infty$) on the line of expectations is not in the C_p reversal region RR for any $\lambda_0 > 0$.*

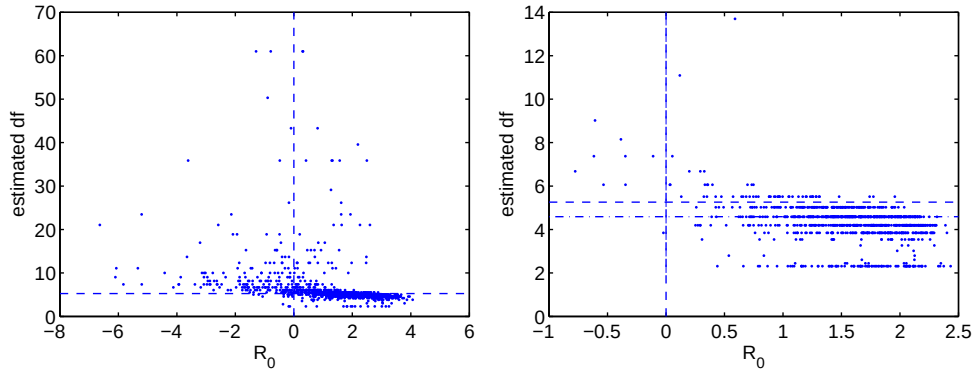
Proof. See the Appendix.

From theorem 1, we have the following conclusions. Since the reversal region is a half space, if $\boldsymbol{u}$ is not in RR, then all points on the line segment $\boldsymbol{u}(\gamma) = \gamma\boldsymbol{u} + (1-\gamma)\mathbf{1}$, $\gamma \in [0,1]$, lie outside RR. Moreover, if $\boldsymbol{u}$ is in RR, then, for all sufficiently small $\gamma \in (0,1)$, the points $\boldsymbol{u}(\gamma)$ lie outside RR. In this case, using (10), we can find the (data dependent) value γ^* such that for all $\gamma < \gamma^*$, $\boldsymbol{u}(\gamma)$ is outside RR. It is defined by

$$\frac{1}{\gamma^*} = 1 - \frac{R_0(\boldsymbol{u})}{R_0(\mathbf{1})} = 1 - \frac{\sum(-\ddot{\eta}_{\lambda_0 i})u_i - 2\ddot{b}_{\lambda_0 i} - \beta_{\lambda_0}\sum(-\dot{\eta}_{\lambda_0 i})u_i - 2\dot{b}_{\lambda_0 i}}{\sum(-\ddot{\eta}_{\lambda_0 i}) - 2\ddot{b}_{\lambda_0 i} - \beta_{\lambda_0}\sum(-\dot{\eta}_{\lambda_0 i}) - 2\dot{b}_{\lambda_0 i}}.$$

For Example 1 of Kou & Efron (2002), simulations reveal that, with $\lambda_0 = \lambda_{ET}$, the (empirical) distribution of γ^* is supported on $[0.2, 1]$ (approximately), and near 0.2 the density approaches 0 continuously. This indicates that RC_p with $\gamma = 0.2$ is very stable.

Since RGCV is related to RC_p in much the same way as GCV is related to C_p, we can expect that RGCV will display the same stability property as RC_p. For the same 1000 replicates used for GCV in Fig. 1(a), the reversal effect for RGCV with $\gamma = 0.5$ is much less serious, with only 5.3% of the replicates having $R_0(\boldsymbol{u}(\gamma)) < 0$. Figure 1(b) shows the corresponding results for RGCV with $\gamma = 0.3$. Now only 0.9% of the replicates have $R_0(\boldsymbol{u}(\gamma)) < 0$, and for all of the 1000 replicates, the RGCV estimated *df* is between 2.3 and 13.7, a big improvement compared to GCV in Fig. 1(a).



(a) (b)

Figure 1: Reversal results for (a) GCV and (b) RGCV ($\gamma = 0.3$) with $\lambda_0 = \lambda_{ET}$ (i.e. $df = 5.26$, marked with a horizontal line) for Ex. 1 of Kou & Efron (2002). The lower horizontal line marks the value $df = 4.59$ corresponding to λ_{EW}, the minimizer of $EW(\lambda)$.

It is clear from Fig. 1(b) that RGCV introduces a downward bias in the estimated *df*. This is to be expected from the asymptotic estimate in corollary 1 in section 3. Some bias in this direction is helpful since, as discussed in the Introduction, the Sobolev risk $EW(\lambda)$ defined in (5) is a better performance measure than $ET(\lambda)$, and its minimizer λ_{EW} usually has a smaller *df* value than λ_{ET}. In particular, for the example above, λ_{EW} gives $df = 4.59$, which is marked in Fig. 1(b) with the lower horizontal dash-dot line. In other examples to be considered below, there is an even bigger difference between the *df* values corresponding to λ_{ET} and λ_{EW}.

For any estimate $\hat{\lambda}$, define the inefficiency with respect to the Sobolev error to be $I_W = W(\hat{\lambda})/\min W(\lambda)$. Plots of I_W (instead of *df*) against R_0 corresponding to Figs. 1(a) and 1(b) show that, for this example, the sign of R_0 is a good predictor of the instability of GCV measured by I_W, and that RGCV with $\gamma = 0.3$ is much more stable than GCV.

While it is natural to have $\lambda_0 = \lambda_{ET}$ in the definition (10) of R_0 for C_p and GCV, it is of interest to know how sensitive $RR = RR(\lambda_0)$ is to the choice of the 'optimal' λ_0. An obvious comparison is to take $\lambda_0 = \lambda_{EW}$, the minimizer of $EW(\lambda)$, in the definition of R_0. Since GCV is biased as an estimator of λ_{EW}, we can expect that the probability $P(\boldsymbol{u} \in RR(\lambda_{EW}))$ will be

larger than $P(\boldsymbol{u} \in RR(\lambda_{ET}))$. For the same 1000 replicates used in Fig. 1, now 43.1% belong to $RR(\lambda_{EW})$ (and 39.8% of 3000 replicates), which is significantly higher than the corresponding value of 19.7% belonging to $RR(\lambda_{ET})$ (in Fig. 1(a)).

The mean $M(R_0)$, variance $V(R_0)$ and skewness $S(R_0)$ of R_0 are

$$\begin{aligned} M(R_0) &= \lambda_0^2 \left[\sum 2\dot{b}_{\lambda_0 i}^2 \mu_{\lambda_0 i} + (\beta_{\lambda_0} \dot{\boldsymbol{\eta}}_{\lambda_0} - \ddot{\boldsymbol{\eta}})^T (E\boldsymbol{u} - \boldsymbol{\mu}_{\lambda_0}) \right] \\ V(R_0) &= \lambda_0^4 \sum (\beta_{\lambda_0} \dot{\eta}_{\lambda_0 i} - \ddot{\eta}_{\lambda_0 i})^2 \mathrm{var}(u_i) \\ S(R_0) &= \lambda_0^6 \sum (\beta_{\lambda_0} \dot{\eta}_{\lambda_0 i} - \ddot{\eta}_{\lambda_0 i})^3 E(u_i - Eu_i)^3 / [V(R_0)]^{3/2}, \end{aligned}$$

where, for normal errors, $Eu_i = g_i^2 + 1$, $\mathrm{var}(u_i) = 4g_i^2 + 2$ and $E(u_i - Eu_i)^3 = 24g_i^2 + 8$. As in Kou & Efron (2002), a three term Edgeworth expansion yields the approximation

$$P(\boldsymbol{u} \in RR) \approx \Phi\left(\frac{-M(R_0)}{\sqrt{V(R_0)}}\right) - \tfrac{1}{6} S(R_0) \left(\frac{M(R_0)^2}{V(R_0)} - 1\right) \phi\left(\frac{-M(R_0)}{\sqrt{V(R_0)}}\right), \tag{12}$$

where ϕ and Φ are the standard normal density and distribution functions, respectively. Note that there is good agreement between the analytic estimate $P(\boldsymbol{u} \in RR(\lambda_{ET})) = 0.19$ (from (12)) and the empirical percentage 19.7% (from Fig. 1(a)), and also between $P(\boldsymbol{u} \in RR(\lambda_{EW})) = 0.39$ (from (12)) and the empirical percentage 39.8% (for 3000 replicates) discussed above.

Using $\boldsymbol{u}(\gamma) = \gamma \boldsymbol{u} + (1-\gamma)\mathbf{1}$ in place of $\boldsymbol{u}$ in (12) and normal errors, we can obtain an analytic estimate of $P(\boldsymbol{u}(\gamma) \in RR(\lambda_0))$ for the RGCV criterion. The parameter γ appears in $M(R_0)$ and $V(R_0)$ through the expressions $Eu_i(\gamma) = \gamma Eu_i + 1 - \gamma$ and $\mathrm{var}(u_i(\gamma)) = \gamma^2 \mathrm{var}(u_i)$, respectively, but it does not appear in $S(R_0)$ since it cancels out.

Because the RGCV estimate is asymptotically larger than λ_{ET} (see (21)), it is appropriate to use a value of λ_0 in $RR(\lambda_0)$ that is larger than λ_{ET}, say between λ_{ET} and λ_{EW}. Figure 2 shows plots of (the estimate) $P(\boldsymbol{u}(\gamma) \in RR(\lambda_0))$, as a function of $\gamma \in [0,1]$, for six different values of λ_0 defined by $\lambda_0 = s\lambda_{EW} + (1-s)\lambda_{ET}$ for $s = 0, 0.2, \ldots, 1$. For $s = 0$, we have $\lambda_0 = \lambda_{ET}$, which is used in Fig. 1, and, for $s = 1$, we have $\lambda_0 = \lambda_{EW}$. Clearly, for $\gamma = 1$ (i.e. GCV), the probability $P(\boldsymbol{u}(\gamma) \in RR(\lambda_0))$ is in the interval $[0.19, 0.39]$. But, as γ decreases, the corresponding probability interval gets closer to 0; in particular, when $\gamma = 0.3$, the interval has decreased to approximately $[0.004, 0.05]$. This means that, with $\gamma = 0.3$, it is very unlikely that RGCV will behave in an unstable manner because of the reversal effect.

The behaviour observed in Fig. 2 can also be seen in other examples. Using $\lambda_0 = \lambda_{ET}$, Fig. 3(a) displays plots of $P(\boldsymbol{u}(\gamma) \in RR(\lambda_0))$ for Examples 1–3 from Craven & Wahba (1979), which involve a unimodal, bimodal and trimodal function f, respectively, with sample sizes of 50 and 100. Note that GCV has significant instability for Examples 1 and 2 (especially Ex. 1), and this instability is only reduced a little by increasing n from 50 to 100. For the same replicates, Fig. 3(b) shows the corresponding plots of $P(\boldsymbol{u}(\gamma) \in RR(\lambda_0))$ using $\lambda_0 = \lambda_{EW}$. In all the cases shown in Figs. 3(a) and 3(b), we can ensure that $P(\boldsymbol{u}(\gamma) \in RR(\lambda_0))$ (with both $\lambda_0 = \lambda_{ET}$ and $\lambda_0 = \lambda_{EW}$) is very small by taking $\gamma = 0.3$.

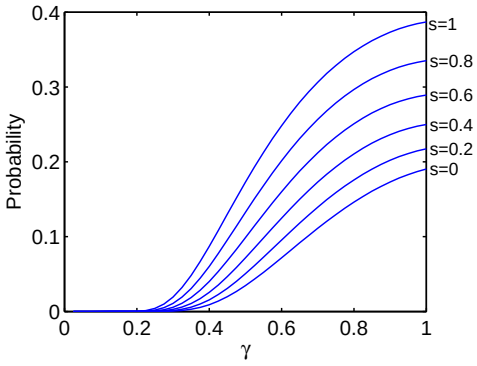
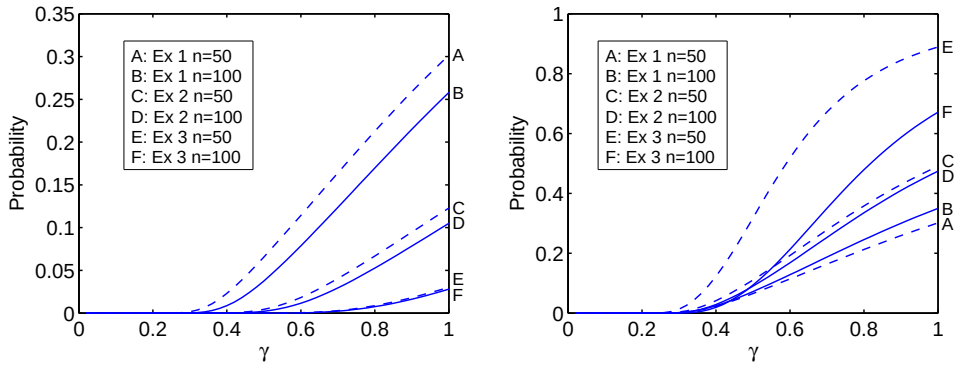


Figure 2: Plots of $P(\boldsymbol{u}(\gamma) \in RR(\lambda_0))$ against γ with $\lambda_0 = s\lambda_{EW} + (1-s)\lambda_{ET}$, $s = 0, 0.2, \ldots, 1$, for Ex. 1 of Kou & Efron (2002)



(a) (b)

Figure 3: Plots of $P(\boldsymbol{u}(\gamma) \in RR(\lambda_0))$ against γ with (a) $\lambda_0 = \lambda_{ET}$ and (b) $\lambda_0 = \lambda_{EW}$ for Ex. 1–3 of Craven & Wahba (1979) with n=50 (dashed) and n=100 (solid)

3 Asymptotic inefficiency of RGCV for the prediction risk

Since the errors are independent with mean 0, the prediction risk can be expressed as $ET(\lambda) = b^2(\lambda) + v(\lambda)$, where $b^2(\lambda) = n^{-1}\|E\boldsymbol{f}_\lambda - \boldsymbol{f}\|^2$ is the squared bias and $v(\lambda) = n^{-1}E\|\boldsymbol{f}_\lambda - E\boldsymbol{f}_\lambda\|^2 = \sigma^2\mu_2(\lambda)$ is the variance. It is known (Lukas, 2008) that, under suitable assumptions, as $n \to \infty$, the shifted RGCV function $\overline{V}(\lambda) - \gamma n^{-1}\|\boldsymbol{\varepsilon}\|^2$ is consistent with the robust prediction risk defined as

$$E\overline{T}(\lambda) = \gamma ET(\lambda) + (1-\gamma)v(\lambda) = \gamma b^2(\lambda) + v(\lambda). \tag{13}$$

Therefore, the RGCV estimate is asymptotically biased for the prediction risk. Here, we determine the asymptotic inefficiency of the (restricted) RGCV estimate for the prediction risk.

We will use the same assumptions as in Cox (1984a); Nychka (1990). Let G_n denote the em-

pirical distribution function for the design points $\{x_i\}$. Cases A and B below cover deterministic and random design points, respectively.

Case A. There is a distribution function G such that $\sup|G_n(x) - G(x)| = O(1/n)$. If $\{x_i\}$ are uniformly spaced, then this holds with $G(x) = (x-a)/(b-a)$.

Case B. The set $\{x_i\}$ is a random sample from a distribution with c.d.f. G.
In both cases, we assume that $G \in C^\infty[a,b]$ and G' is strictly positive on $[a,b]$.

Assumption A1. The random errors ε_i in (1) satisfy $E|\varepsilon_i|^{2+\nu} < \infty$, with $\nu > 4m-1$ (case A), and $\nu > 2(8m-3)/5$ (case B).

As for GCV in Nychka (1990), the convergence results will apply to the RGCV estimate restricted to $\lambda \geq \alpha_n$ for a positive sequence $\alpha_n \to 0$ as $n \to \infty$.

Assumption A2. As $n \to \infty$, the sequence $\alpha_n \to 0$ at the rate $\alpha_n \approx n^{-4m/5}\log(n)$ (case A), and $\alpha_n \approx n^{-2m/5}(\log(n))^m$ (case B). Here $\alpha_n \approx \beta_n$ means that $c_1\beta_n \leq \alpha_n \leq c_2\beta_n$ for some positive constants c_1 and c_2.

Assumption A3. There are constants $p \in [1,2]$ and $c = c(p) > 0$ such that as $n \to \infty$, the squared bias satisfies

$$b^2(\lambda) \;=\; n^{-1}\|(I-A)\boldsymbol{f}\|^2 \;=\; c\lambda^p(1+o(1)), \tag{14}$$

uniformly for $\lambda \in [\alpha_n, \infty)$.

It is known that, under certain conditions, assumption A3 holds with p directly related to the smoothness of f, including its boundary behaviour (Cox, 1988; Nychka, 1990). For $f \in \mathcal{W} = \mathcal{W}^{m,2}[a,b]$, we have the bound $b^2(\lambda) \leq c\lambda(1+o(1))$ with $c = \int_a^b [f^{(m)}(x)]^2 dx$. Equality holds for a class of functions in $\mathcal{W}$ for which m is the maximum order of smoothness (i.e. they do not belong to a Sobolev space of higher order), so $p = 1$ for this class. If f belongs to $\mathcal{W}^{2m,2}[a,b]$ and satisfies the natural boundary conditions $f^{(j)}(a) = f^{(j)}(b) = 0$, $j = m, \ldots, 2m-1$, then (14) holds with $p = 2$ and $c = \int_a^b [f^{(2m)}(x)]^2 dx$. Moreover, $p = 2$ is the highest exponent possible, regardless of any higher smoothness of f.

The functions $\mu_1(\lambda) = n^{-1}\mathrm{tr}A(\lambda)$ and $\mu_2(\lambda) = n^{-1}\mathrm{tr}(A^2(\lambda))$ can be estimated using the eigenproblem defining the Demmler–Reinsch basis of natural polynomial splines. From results about the asymptotic behaviour of the eigenvalues (Cox, 1984a; Speckman, 1985) (see also Nychka (1990, lemma 3.1) and Eubank (1988, sect. 6.3.2)), it is known that under assumptions A1 and A2, as $n \to \infty$,

$$\mu_k(\lambda) \;=\; \alpha l_k n^{-1}\lambda^{-1/(2m)}(1+o(1)), \quad k = 1,2, \tag{15}$$

where $\alpha = \pi^{-1}\int_a^b (G'(x))^{1/(2m)}\,dx$ and

$$l_k = \int_0^\infty (1+x^{2m})^{-k}\,dx = \Gamma(1/(2m))\Gamma(k-1/(2m))/(2m\Gamma(k)), \tag{16}$$

uniformly for $\lambda \in [\alpha_n, \infty)$. (Note there is an error in the definition of α in Nychka (1990).) For $m = 2$, we have $l_1 = 10/9$ and $l_2 = 5/6$.

From assumption A3 and the estimate of $\mu_2(\lambda)$ in (15), it follows that

$$ET(\lambda) \;=\; (c\lambda^p + \sigma^2\alpha l_2 n^{-1}\lambda^{-1/(2m)})(1+o(1)), \tag{17}$$

uniformly for $\lambda \in [\alpha_n, \infty)$. Let λ_{ET} be the minimizer of $ET(\lambda)$ for $\lambda \geq \alpha_n$. Minimizing the right-hand side in (17) gives the known estimate (Nychka, 1990; Wahba, 1990)

$$\lambda_{ET} \;=\; \left(\frac{\alpha l_2 \sigma^2}{2mpcn}\right)^{2m/(2mp+1)} (1+o(1)) \tag{18}$$

(which is in $[\alpha_n, \infty)$ for all sufficiently large n).

Define $S^2 = n^{-1}\|\boldsymbol{\varepsilon}\|^2$. It is known (Nychka, 1990, lemma 3.1) that, as $n \to \infty$, the shifted GCV function $V(\lambda) - S^2$ is consistent with the prediction error $T(\lambda)$ and risk $ET(\lambda)$, and it follows (Nychka, 1990, lemma 3.2) that if $\hat{\lambda}_V$ minimizes $V(\lambda)$ for $\lambda \geq \alpha_n$, then $\hat{\lambda}_V = \lambda_{ET}(1 + o_P(1))$. These results can be extended to RGCV by showing that the shifted RGCV function $\overline{V}(\lambda) - \gamma S^2$ is consistent with the robustified prediction error $\overline{T}(\lambda) = \gamma T(\lambda) + (1-\gamma)v(\lambda)$. Using the same argument as in theorem 4.1 of Lukas (2008), we have the following result.

Theorem 2 *Under assumptions A1–A3, as $n \to \infty$,*

$$\sup_{\lambda\in[\alpha_n,\infty)} \left| \frac{\overline{V}(\lambda) - \gamma S^2 - \overline{T}(\lambda)}{E\overline{T}(\lambda)} \right| = o_P(1), \quad \sup_{\lambda\in[\alpha_n,\infty)} \left| \frac{\overline{T}(\lambda) - E\overline{T}(\lambda)}{E\overline{T}(\lambda)} \right| = o_P(1) \tag{19}$$

and

$$E\overline{T}(\lambda) = (\gamma c\lambda^p + \sigma^2\alpha l_2 n^{-1}\lambda^{-1/(2m)})(1+o(1)), \tag{20}$$

uniformly for $\lambda \in [\alpha_n, \infty)$.

Define $\hat{\lambda}_{\overline{V}}$ to be the minimizer of the RGCV function $\overline{V}(\lambda)$ for $\lambda \geq \alpha_n$.

Corollary 1 *The* RGCV *estimate $\hat{\lambda}_{\overline{V}}$ satisfies*

$$\hat{\lambda}_{\overline{V}} \;=\; \left(\frac{\alpha l_2 \sigma^2}{2mp\gamma cn}\right)^{\frac{2m}{2mp+1}} (1+o_P(1)) \;=\; \gamma^{-\frac{2m}{2mp+1}} \lambda_{ET}(1+o_P(1)) \tag{21}$$

as $n \to \infty$.

Proof. Comparing (20) and (17), it is clear that the minimizer $\lambda_{E\overline{T}}$ of $E\overline{T}(\lambda)$ for $\lambda \geq \alpha_n$ is the same as that in (18) with c replaced by γc. Then, the estimate of $\hat{\lambda}_{\overline{V}}$ in (21) follows from theorem 2 using the same argument as in lemma 3.2 of Nychka (1990).

Corollary 1 shows that $\hat{\lambda}_{\overline{V}}$ has the same optimal decay rate as λ_{ET}. But, since $0 < \gamma < 1$, $\hat{\lambda}_{\overline{V}}$ is asymptotically larger than λ_{ET}, as we would expect. With $m = 2$, $p = 2$ and $\gamma = 0.5$, the factor is $\gamma^{-2m/(2mp+1)} = 1.36$.

The following result gives the asymptotic inefficiency of $\hat{\lambda}_{\overline{V}}$ for the prediction error and risk.

Corollary 2 *Suppose that $\hat{\lambda}_T$ minimizes $T(\lambda)$ for $\lambda \geq \alpha_n$. Under assumptions A1–A3, as $n \to \infty$,*

$$I_{ET} = \frac{ET(\hat{\lambda}_{\overline{V}})}{ET(\lambda_{ET})} = K(1+o_P(1)),\; I_T = \frac{T(\hat{\lambda}_{\overline{V}})}{T(\hat{\lambda}_T)} = K(1+o_P(1)), \tag{22}$$

where

$$K = \frac{(2mp\gamma)^{-2mp/(2mp+1)} + (2mp\gamma)^{1/(2mp+1)}}{(2mp)^{-2mp/(2mp+1)} + (2mp)^{1/(2mp+1)}}. \tag{23}$$

Proof. The first equality in (22) is found by substituting the estimates obtained for $\hat{\lambda}_{\overline{V}}$ and λ_{ET} into the estimate for $ET(\lambda)$ in (17). The second equality follows from the first equality since, from (19) (with $\gamma = 1$) and because $ET(\lambda_{ET}) \leq ET(\hat{\lambda}_T)$, we have

$$1 \leq \frac{T(\hat{\lambda}_{\overline{V}})}{T(\hat{\lambda}_T)} = \frac{T(\hat{\lambda}_{\overline{V}})}{ET(\hat{\lambda}_{\overline{V}})} \frac{ET(\hat{\lambda}_{\overline{V}})}{ET(\lambda_{ET})} \frac{ET(\lambda_{ET})}{ET(\hat{\lambda}_T)} \frac{ET(\hat{\lambda}_T)}{T(\hat{\lambda}_T)} = \frac{ET(\hat{\lambda}_{\overline{V}})}{ET(\lambda_{ET})}(1 + o_P(1)).$$

This completes the proof.

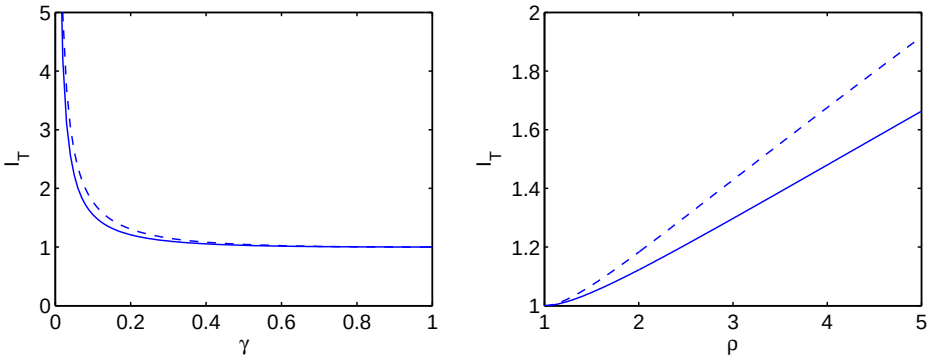
It is easy to show that, for any m and p, the value K of the asymptotic inefficiency I_T in (23) is a strictly decreasing function of γ, with $K = 1$ at $\gamma = 1$ and $K \to \infty$ as $\gamma \to 0$. The value $K = 1$ at $\gamma = 1$ reflects the fact that GCV is asymptotically optimal for the prediction risk, and the monotonic nature of the function is consistent with the fact that, as γ decreases, RGCV becomes increasingly biased in estimating λ_{ET}. Fig. 4(a) shows the graph of K as a function of γ for $m = 2$ (cubic splines) with $p = 1$ (dashed) and $p = 2$ (solid). Table 1 shows the corresponding values of K for several values of γ. Clearly, from the shape of the graphs, there is a large interval of γ values for RGCV for which the asymptotic inefficiency I_T is close to 1, in fact $I_T \leq 1.153$ for all $\gamma \in [0.3, 1]$. Consequently, we can expect that, for any $\gamma \in [0.3, 1]$, RGCV will perform well for large n.

Note that, since the expression for K in (23) does not involve α or c, the asymptotic inefficiency I_T of RGCV is independent of the scale of the interval $[a, b]$.

It is shown in theorem 4.2 of Lukas (2008) that, under certain conditions, the modified GCV criterion, with score function $V_\rho(\lambda)$ in (6), is asymptotically equivalent to RGCV. The result applies directly here under assumptions A1–A3, since it follows easily from (17) that $nET(\lambda) \to \infty$ as $n \to \infty$ for $\lambda \geq \alpha_n$, and, if $\mu_1(\lambda) \to 0$, then, from (15), we have $\mu_1(\lambda)/\mu_2(\lambda) \to l_1/l_2$. Hence, if $\mu_1(\lambda) \to 0$ and $\gamma^{-1} = 1 + 2(\rho - 1)l_1/l_2$, then $\gamma^{-1}\overline{V}(\lambda) - V_\rho(\lambda) = o_P(\overline{R}(\lambda))$. Therefore, from theorem 2 and corollary 1, we obtain the following corollary for modified GCV.

Corollary 3 *Suppose that $\hat{\lambda}_{V_\rho}$ minimizes $V_\rho(\lambda)$ for $\lambda \geq \alpha_n$. Under assumptions A1–A3, $\hat{\lambda}_{V_\rho}$ satisfies $\hat{\lambda}_{V_\rho} = \gamma^{-2m/(2mp+1)}\lambda_{ET}(1 + o_P(1))$ and has the same asymptotic inefficiency I_{ET} and I_T as for $\hat{\lambda}_{\overline{V}}$ in corollary 2, where $\gamma^{-1} = 1 + 2(\rho - 1)l_1/l_2$.*

Using this result for cubic splines (i.e. $m = 2$, for which $l_1/l_2 = (10/9)/(5/6) = 4/3$), we obtain the plot in Fig. 4(b) of the asymptotic inefficiency I_T against ρ for modified GCV. Some corresponding values are given in Table 1. Consequently, we can expect that, for any $\rho \in [1, 1.875]$, modified GCV will perform well for large n.



(a) (b)

Figure 4: Asymptotic inefficiency I_T as a function of (a) γ for RGCV and (b) ρ for modified GCV, with $m = 2$, $p = 1$ (dashed) and $p = 2$ (solid)

Table 1: Values of asymptotic inefficiencies I_T for RGCV and modified GCV with $m = 2$

γ	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1
ρ	8.125	4.375	2.5	1.875	1.5625	1.375	1.25	1.0938	1
I_T $(p = 1)$	2.636	1.767	1.305	1.153	1.082	1.045	1.023	1.0042	1
I_T $(p = 2)$	2.230	1.548	1.208	1.102	1.054	1.029	1.015	1.0026	1

4 Asymptotic inefficiency of RGCV for the Sobolev risk

The prediction error $T(\lambda)$ has limitations as a measure of the quality of the fit of a spline estimate f_λ. It is only a pointwise measure and, furthermore, it is insensitive to discrepancies in the slope and curvature of f_λ, which are important for the quality of the fit. To see this, consider the prediction error as an approximation of the squared $L_2(G)$ norm error $\int_a^b h^2 dG$, where $h = f_\lambda - f$. If $G(x) = x$ (uniform points) and $h(x) = c\sin(k\pi x)$, then $\int_0^1 (h(x))^2 dx$ is independent of k, while the integrated squared (linear) curvature $\int_0^1 (h''(x))^2 dx$ is proportional to k^4. In this situation, even though the prediction error will be small if c is small, f_λ would be judged to be too rough if k is large.

When assessing the accuracy of f_λ compared to f by eye, one intuitively takes into account not only function values but also the slope and curvature. This suggests the use of a continuous error involving integrated squares of $f_\lambda(x) - f(x)$ and its first and some higher derivatives. However, it is only necessary to include $f_\lambda(x) - f(x)$ and the highest of the derivatives, since then the error will automatically be sensitive to discrepancies in lower order derivatives. This follows from the fact (see theorem 2.5 in Schumaker (1981)) that, for each integer $j < J$, there is a constant C

such that

$$\int_a^b (h^{(j)})^2(x)dx \le C\left[\int_a^b h^2(x)dx + \int_a^b (h^{(J)})^2(x)dx\right]$$

for all h. Thus, since f_λ is defined by (2) using the mth derivative and f is assumed to belong to $W^{m,2}[a,b]$, it is natural and reasonable to use the Sobolev error $W(\lambda)$ defined in (5). The case with $m = 2$ is the most common.

The asymptotic behaviour of the Sobolev risk $EW(\lambda)$ was studied in Cox (1984b, 1988); Wahba & Wang (1990); Lukas (1993). It is known (Cox, 1984b, theorem 5.1) that, under suitable assumptions, if $f \in W^{q,2}[a,b]$ for $m < q \le 3m$, then

$$EW(\lambda) = \lambda^{-1}O(\lambda^{q/m} + n^{-1}\lambda^{-1/(2m)}),$$

uniformly for λ in a certain interval depending on n. Therefore, $EW(\lambda)$ has best possible rate if $\lambda = \lambda_{EW} \approx n^{-2m/(2q+1)}$. Using the appropriate substitution $q = mp$, it can be seen that this optimal rate for λ_{EW} is the same as the optimal rate for λ_{ET} in (18) for the prediction risk, but, as we shall see, the constants are different. To find the asymptotic inefficiency of the RGCV estimate, we will estimate $EW(\lambda)$ more precisely.

First we define an error function that approximates the Sobolev error and is easier to estimate. Let $f_{\rm int}$ be the natural polynomial spline of degree $2m-1$ that interpolates $f(x)$ at the points x_i, $i = 1,\ldots,n$. It is well known (de Boor & Lynch, 1966) that $f_{\rm int}$ is the unique minimizer of $\int_a^b[\phi^{(m)}(x)]^2dx$ subject to $\phi(x_i) = f(x_i)$, $i = 1,\ldots,n$. Let $\mathcal{S}$ denote the n dimensional vector space of natural spline functions of degree $2m-1$ with knots at x_i, $i = 1,\ldots,n$. Define the Hilbert space $\widetilde{\mathcal{W}}$ to be the set $\mathcal{W}^{m,2}[a,b]$ with the inner product

$$(f,g)_{\widetilde{\mathcal{W}}} \;=\; n^{-1}\sum_{i=1}^n f(x_i)g(x_i) + \int_a^b f^{(m)}(x)g^{(m)}(x)\,dx. \tag{24}$$

Let $P_{\mathcal{S}}$ be the orthogonal projection of $\widetilde{\mathcal{W}}$ onto $\mathcal{S}$.

Lemma 1 *For any $f \in \widetilde{\mathcal{W}}$, we have $P_{\mathcal{S}}f = f_{\rm int}$.*

Proof. Using the definition of $\widetilde{\mathcal{W}}$, we obtain

$$\|f - f_{\rm int}\|^2_{\widetilde{\mathcal{W}}} = \|f^{(m)} - f^{(m)}_{\rm int}\|^2_{L_2} \le \|f^{(m)} - \phi^{(m)}\|^2_{L_2} \le \|f-\phi\|^2_{\widetilde{\mathcal{W}}}$$

for any $\phi \in \mathcal{S}$, where the second last inequality is a well known minimum property of $f_{\rm int}$ (de Boor & Lynch, 1966). The result follows.

Lemma 1 shows that $f_{\rm int}$ is the best approximation of f from $\mathcal{S}$ in the sense that it minimizes $\|f-\phi\|^2_{\widetilde{\mathcal{W}}}$ for $\phi \in \mathcal{S}$. Moreover, since $f_\lambda \in \mathcal{S}$, we have

$$\widetilde{W}(\lambda) \;:=\; \|f_\lambda - f\|^2_{\widetilde{\mathcal{W}}} \;=\; \|f_\lambda - f_{\rm int}\|^2_{\widetilde{\mathcal{W}}} + \|f_{\rm int} - f\|^2_{\widetilde{\mathcal{W}}}. \tag{25}$$

Since $\|f_{\rm int} - f\|^2_{\widetilde{\mathcal{W}}}$ is independent of λ, the error $\widetilde{W}(\lambda)$ and $\|f_\lambda - f_{\rm int}\|^2_{\widetilde{\mathcal{W}}}$ have the same minimizer. Similarly, $E\widetilde{W}(\lambda)$ and $E\|f_\lambda - f_{\rm int}\|^2_{\widetilde{\mathcal{W}}}$ have the same minimizer $\lambda_{E\widetilde{W}}$. This and (25) imply that,

for any λ,

$$\frac{E\widetilde{W}(\lambda)}{\min_\lambda E\widetilde{W}(\lambda)} = \frac{E\|f_\lambda - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}} + \delta}{E\|f_{\lambda_{E\widetilde{W}}} - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}} + \delta} \leq \frac{E\|f_\lambda - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}}}{\min_\lambda E\|f_\lambda - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}}}, \tag{26}$$

where $\delta = \|f_{\text{int}} - f\|^2_{\widetilde{\mathcal{W}}} = \|f^{(m)}_{\text{int}} - f^{(m)}\|^2_{L_2}$. Consequently, we can use the right-hand side of (26) to bound the inefficiency on the left-hand side. The bound will be close if δ is relatively small. This will be the case if f is sufficiently smooth and satisfies the same boundary conditions as f_{int}; in particular, if $f \in \mathcal{W}^{m+1,2}[a,b]$ and the points x_i are equally spaced, then $\delta = O(n^{-2})$ (Swartz & Varga, 1972), which is a much faster rate than that of $E\|f_{\lambda_{E\widetilde{W}}} - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}}$ as $n \to \infty$ (see (38) and (39)). Since the errors are independent with mean 0, we have

$$\begin{aligned} E\|f_\lambda - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}} &= En^{-1}\|\boldsymbol{f}_\lambda - \boldsymbol{f}\|^2 + E\|(f_\lambda - f_{\text{int}})^{(m)}\|^2_{L_2} \\ &= b^2(\lambda) + v(\lambda) + b_1^2(\lambda) + v_1(\lambda), \end{aligned} \tag{27}$$

where $b_1^2(\lambda) = \|Ef^{(m)}_\lambda - f^{(m)}_{\text{int}}\|^2_{L_2}$ and $v_1(\lambda) = E\|f^{(m)}_\lambda - Ef^{(m)}_\lambda\|^2_{L_2}$.

Consider the same diagonalization of the smoothing matrix $A(\lambda)$ as in section 2, i.e. $A(\lambda) = U\text{diag}(a_{\lambda i})U^T$, where U is orthogonal and $a_{\lambda i} = 1/(1+\lambda\tau_i)$, $i = 1,\ldots,n$, for a certain nondecreasing sequence $\{\tau_i\}$, with $\tau_i = 0$, $i = 1,\ldots,m$. Let $\mu_1(\lambda) = n^{-1}\text{tr}A(\lambda)$, $\mu_2(\lambda) = n^{-1}\text{tr}(A^2(\lambda))$ (as above) and $\mu_{12}(\lambda) = (\mu_1(\lambda) - \mu_2(\lambda))/\lambda$.

Lemma 2 *If the errors ε_i are independent with mean 0 and variance σ^2, then $v(\lambda) = \sigma^2\mu_2(\lambda)$, $v_1(\lambda) = \sigma^2\mu_{12}(\lambda)$,*

$$\begin{aligned} b^2(\lambda) &= \lambda^2 n^{-1}\sum(\tau_i^{-1} + \lambda)^{-2}(U^T\boldsymbol{f})_i^2 \quad \textit{and} \\ b_1^2(\lambda) &= \lambda^2 n^{-1}\sum\tau_i(\tau_i^{-1} + \lambda)^{-2}(U^T\boldsymbol{f})_i^2. \end{aligned}$$

Proof. See the Appendix.

Analogous to assumption A3, it will be assumed that:

Assumption A4. There are constants $p \in (1,2)$, c and c_1 such that, as $n \to \infty$,

$$b^2(\lambda) = c\lambda^p(1 + o(1)) \quad \text{and} \quad b_1^2(\lambda) = c_1\lambda^{p-1}(1 + o(1)),$$

uniformly for $\lambda \in [\alpha_n, \infty)$.

For assumption A4 to hold, it is necessary (Cox, 1988; Lukas, 1993) that f has smoothness between that corresponding to $\mathcal{W}^{m,2}[a,b]$ and $\mathcal{W}^{2m,2}[a,b]$.

Definition 1 *Let $\mathcal{C}$ denote the class of problems where $m = 2$, the points x_i are equally spaced and f satisfies $n^{-1}(U^T\boldsymbol{f})_i^2 = k\tau_i^{-r}$ for constants $k > 0$ and $r \in (5/4, 9/4)$.*

Lemma 3 *For problems in class $\mathcal{C}$, if $n^4\alpha_n \to \infty$, then assumption A4 holds with $p = r - 1/4$.*

Proof. See the Appendix.

Define $\mathcal{W}$ to be the Sobolev space $\mathcal{W}^{m,2}[a,b]$ with inner product

$$(f,g)_{\mathcal{W}} = \int_a^b f(x)g(x)dG + \int_a^b f^{(m)}(x)g^{(m)}(x)dx$$

and let $W(\lambda) = \|f_\lambda - f\|^2_{\mathcal{W}}$ as in (5). Clearly, $\widetilde{W}(\lambda) - W(\lambda)$ is equal to the error in the discrete approximation $n^{-1}\sum(f_\lambda - f)^2(x_i)$ of $\int(f_\lambda - f)^2 dG$. Using the assumption of independent errors with mean 0, we have

$$E\widetilde{W}(\lambda) - EW(\lambda) = b^2(\lambda) - b_G^2(\lambda) + v(\lambda) - v_G(\lambda),$$

where $b_G^2(\lambda) = \|Ef_\lambda - f\|^2_{L_2(G)}$ and $v_G(\lambda) = E\|f_\lambda - Ef_\lambda\|^2_{L_2(G)}$. It is known (Cox, 1984a; Lukas, 1993) that, under certain assumptions, $v(\lambda) = v_G(\lambda)(1+o(1))$ as $n \to \infty$. In addition, if $b_G^2(\lambda) \approx \min\{1, \lambda^p\}$ for $p \le 2$, then $b^2(\lambda) = b_G^2(\lambda)(1+o(1))$. In particular, these estimates hold for class $\mathcal{C}$. Therefore, since $b_G^2(\lambda) \le EW(\lambda)$ and $v_G(\lambda) \le EW(\lambda)$, it is reasonable to make the following assumption.

Assumption A5. As $n \to \infty$, $E\widetilde{W}(\lambda) = EW(\lambda)(1+o(1))$, uniformly for all $\lambda \in [\alpha_n, \infty)$.

We can now estimate the Sobolev risk $EW(\lambda)$ and the weighted Sobolev risk $EW_\kappa(\lambda) = E\int(f_\lambda - f)^2 dG + \kappa E\int(f_\lambda^{(m)} - f^{(m)})^2 dx$ for the (restricted) RGCV estimate.

Theorem 3 *Suppose that $\hat{\lambda}_{\overline{V}}$ and λ_{EW} minimize $\overline{V}(\lambda)$ and $EW(\lambda)$, respectively, for $\lambda \ge \alpha_n$. Under assumptions A1, A2, A4 and A5, as $n \to \infty$, we have*

$$I_{EW}(\hat{\lambda}_{\overline{V}}) = \frac{EW(\hat{\lambda}_{\overline{V}})}{EW(\lambda_{EW})} \le K_1(1 + o_P(1)), \tag{28}$$

where

$$K_1 = \frac{(w\gamma/\gamma^*)^{-2m(p-1)/(2mp+1)} + (w\gamma/\gamma^*)^{(2m+1)/(2mp+1)}}{w^{-2m(p-1)/(2mp+1)} + w^{(2m+1)/(2mp+1)}} \tag{29}$$

with

$$w = \frac{2m(p-1)}{2m+1} \quad \text{and} \quad \gamma^* = \frac{c_1 l_2 (p-1)}{cp(l_1 - l_2)(2m+1)}. \tag{30}$$

The bound $K_1 = K_1(\gamma)$, for $\gamma \in [0,1]$, has minimum value 1 at $\gamma = \gamma^*$ if $\gamma^* \le 1$. If, in addition, $\|f_{\text{int}} - f\|^2_{\widetilde{\mathcal{W}}} = o(\lambda^{p-1} + n^{-1}\lambda^{-1-1/(2m)})$ uniformly for all $\lambda \ge \alpha_n$, then

$$\lambda_{EW} = \left(\frac{\alpha(l_1 - l_2)(2m+1)\sigma^2}{2mc_1(p-1)n}\right)^{2m/(2mp+1)} (1 + o(1)) \tag{31}$$

and $I_{EW}(\hat{\lambda}_{\overline{V}}) = K_1(1 + o_P(1))$. For class $\mathcal{C}$ (where $m = 2$), we have $\gamma^* = 0.6(2-p)/p \in (0, 0.6)$ for $p \in (1,2)$. The same result holds for the inefficiency with respect to the weighted Sobolev risk $EW_\kappa(\lambda)$, independent of κ.

Proof. See the Appendix.

From (31) and (18), the ratio $\lambda_{EW}/\lambda_{ET}$ can be evaluated for class $\mathcal{C}$ (using $l_1/l_2 = 4/3$ and $c_1/c = (2-p)/(p-1)$ from (40)) to obtain

$$\lim_{n\to\infty} \lambda_{EW}/\lambda_{ET} = (p/(0.6(2-p))^{4/(4p+1)}. \tag{32}$$

This is a strictly increasing function of $p \in (1,2)$, with $\lim \lambda_{EW}/\lambda_{ET} = 1.505$ at $p = 1$ and $\lim \lambda_{EW}/\lambda_{ET} \to \infty$ as $p \to 2$. Therefore, for large n, $\lambda_{EW} > 1.5\lambda_{ET}$ for all $p \in (1,2)$.

For class $\mathcal{C}$, Fig. 5(a) shows the graph of the estimate K_1 of the asymptotic inefficiency $I_{EW}(\hat{\lambda}_{\overline{V}})$ as a function of γ for the three values of $p = 1.1, 1.5, 1.9$. Note that, unlike the monotonic graphs of I_T in Fig. 4(a), the graphs in Fig. 5(a) have both decreasing and increasing sections. In all cases, there is an initial improvement in the efficiency of the RGCV estimate as γ decreases from 1 (i.e. GCV). This reflects the fact that GCV is biased in estimating λ_{EW}, while, for γ near 1, the RGCV estimate is asymptotically larger than λ_{ET} (see (21)) and hence closer to λ_{EW}. The improvement is greatest for the smoothest case, i.e. $p = 1.9$, for which $\lim \lambda_{EW}/\lambda_{ET}$ is largest (equal to 4.988 from (32)).

If γ is decreased too far, the RGCV estimate is too biased and the Sobolev risk grows. In fact $K_1 \to \infty$ as $\gamma \to 0$, though not as quickly as $K \to \infty$ from (23), since $2m(p-1) < 2mp$. The minimum value of K_1 is 1 at $\gamma = \gamma^*$, and therefore, from (28), the asymptotic inefficiency $I_{EW(\hat{\lambda}_{\overline{V}})}$ is also 1 for $\gamma = \gamma^*$. When $\gamma = \gamma*$, from (21) and (31), the RGCV estimate satisfies $\hat{\lambda}_{\overline{V}} = \lambda_{EW}(1 + o_P(1))$.

The optimal value $\gamma^* = 0.6(2-p)/p$ decreases as the smoothness of f (and hence p) increases. When $p = 1.5$, the optimal value is $\gamma^* = 0.2$. Clearly, from Figs. 4(a) and 5(a), for p in a large subinterval of $(1,2)$, RGCV has good large-sample performance for any $\gamma \in [0.2, 0.4]$. Moreover, the results in section 2 indicate that, for $\gamma \in [0.2, 0.4]$, RGCV has strong small-sample stability. A large simulation study in Lukas *et al.* (2008) confirms that RGCV performs well for $\gamma \in [0.2, 0.4]$.

From the asymptotic equivalence of RGCV and the modified GCV criterion (see corollary 3), we obtain the following corollary.

Corollary 4 *Suppose that $\hat{\lambda}_{V_\rho}$ minimizes $V_\rho(\lambda)$ for $\lambda \geq \alpha_n$. Under assumptions A1, A2, A4 and A5, the conclusions of theorem 3 also hold for $\hat{\lambda}_{V_\rho}$, with $\gamma^{-1} = 1 + 2(\rho - 1)l_1/l_2$.*

For class $\mathcal{C}$, Fig. 5(b) shows the graph of K_1 as a function of ρ for modified GCV. From this and the plot of the asymptotic inefficiency I_T in Fig. 4(b), it is clear that, for p in a large subinterval of $(1,2)$, the modified GCV criterion has good large-sample performance for any $\rho \in [1.5, 2.5]$.

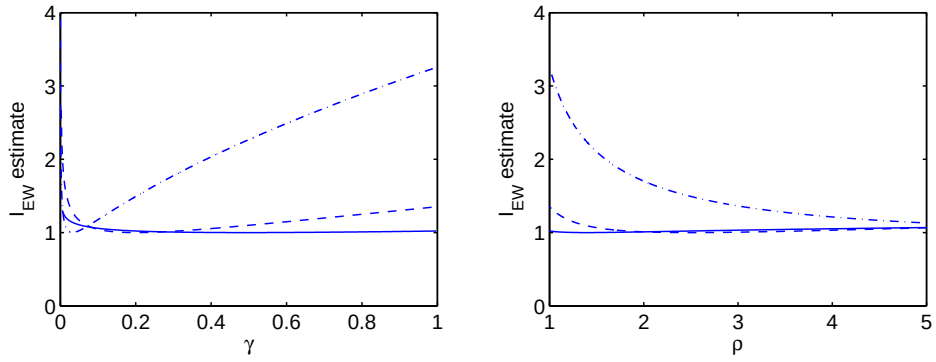
References

Claeskens, G., Krivobokova, T. & Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika* **96**, 529–544.

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

Cox, D. D. (1984b). Multivariate smoothing spline functions. *SIAM J. Numer. Anal.* **21**, 789–813.

Cox, D. D. (1988). Approximation of method of regularization estimators. *Ann. Statist.* **16**, 694–712.



(a) (b)

Figure 5: Asymptotic inefficiency I_{EW} estimate as a function of (a) γ for RGCV and (b) ρ for modified GCV, with $m = 2$, $p = 1.1$ (solid), $p = 1.5$ (dashed) and $p = 1.9$ (dash-dot)

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, D. J., Filloon, T. G. & Nychka, D. (2001). Confidence intervals for nonparametric curve estimates: Toward more uniform pointwise coverage. *J. Amer. Statist. Assoc.* **96**, 233–246.

de Boor, C. & Lynch, R. E. (1966). On splines and their minimum properties. *J. Math. Mech.* **15**, 953–969.

Efron, B. (2001). Selection criteria for scatterplot smoothers. *Ann. Statist.* **29**, 470–504.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression.* Dekker, New York.

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

Hall, P. & Robinson, A. P. (2009). Reducing variability of crossvalidation for smoothing parameter choice. *Biometrika* **96**, 175–186.

Kim, Y.-J. & Gu, C. (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *J. Roy. Statist. Soc. Ser. B* **66**, 337–356.

Kimeldorf, G. S. & Wahba, G. (1971). Some results on Tchebycheffian spline functions and stochastic processes. *J. Math. Anal. Appl.* **33**, 82–95.

Kou, S. C. (2003). On the efficiency of selection criteria in spline regression. *Probab. Theory Related Fields* **127**, 153–176.

Kou, S. C. & Efron, B. (2002). Smoothers and the C_p, generalized maximum likelihood, and extended exponential criteria: a geometric approach. *J. Amer. Statist. Assoc.* **97**, 766–782.

Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14**, 1101–1112.

Lukas, M. A. (1993). Asymptotic optimality of generalized cross-validation for choosing the regularization parameter. *Numer. Math.* **66**, 41–66.

Lukas, M. A. (2006). Robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems* **22**, 1883–1902.

Lukas, M. A. (2008). Strong robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems* **24**, 034006.

Lukas, M. A., de Hoog, F. R. & Anderssen, R. S. (2008). Spline smoothing using robust GCV. Tech. Rep. 08-154, CMIS, CSIRO.

Lukas, M. A., de Hoog, F. R. & Anderssen, R. S. (2010). Efficient algorithms for robust generalized cross-validation spline smoothing. *J. Comput. Appl. Math.* **235**, 102–107.

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

Mitrinović, D. S., Pečarić, J. E. & Fink, A. M. (1993). *Classical and new inequalities in analysis*. Kluwer, Dordrecht.

Nychka, D. (1990). The average posterior variance of a smoothing spline and a consistent estimate of the average squared error. *Ann. Statist.* **18**, 415–428.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional data analysis*. Springer, New York.

Robinson, T. & Moyeed, R. (1989). Making robust the cross-validatory choice of smoothing parameter in spline smoothing regression. *Comm. Statist. Theory Methods* **18**, 523–539.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric regression.* Cambridge University Press, Cambridge.

Schumaker, L. L. (1981). *Spline functions: Basic theory*. Wiley, New York.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13**, 970–983.

Swartz, B. K. & Varga, R. S. (1972). Error bounds for spline and L-spline interpolation. *J. Approx. Theory* **6**, 6–49.

van der Linde, A. (2000). Variance estimation and smoothing-parameter selection for spline regression. *In Smoothing and regression. Approaches, computation and application* (ed M. G. Schimek), 19–41. Wiley, Chichester.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378–1402.

Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadelphia.

Wahba, G. & Wang, Y. (1990). When is the optimal regularization parameter insensitive to the choice of the loss function? *Comm. Statist. Theory Methods* **19**, 1685–1700.

Mark A. Lukas, Mathematics and Statistics, Murdoch University, South Street, Murdoch WA 6150, Australia.
E-mail: M.Lukas@murdoch.edu.au

5 Appendix

Proof of theorem 1
Some simple algebra gives

$$\begin{aligned}\dot{l}_\lambda(\mathbf{1}) &= -2\sum \dot{b}_{\lambda i}(1-b_{\lambda i}) = -2\sum \tau_i(1+\lambda\tau_i)^{-3},\\ \ddot{l}_\lambda(\mathbf{1}) &= 2\sum \dot{b}_{\lambda i}^2 - \ddot{b}_{\lambda i}(1-b_{\lambda i}) = 6\sum \tau_i^2(1+\lambda\tau_i)^{-4},\end{aligned}$$

where, for simplicity, we use λ in place of λ_0. The sums are over i with $\tau_i > 0$. Using the expression for β_λ in the appendix of Kou & Efron (2002), we have

$$\begin{aligned}\beta_\lambda &= -\lambda^{-1}\left[2 - 3(\textstyle\sum a_{\lambda i}^3 b_{\lambda i}^{-2})/(\sum a_{\lambda i}^2 b_{\lambda i}^{-2})\right]\\ &= -\lambda^{-1}\left[2 - \textstyle\sum 3\tau_i^{-2}(1+\lambda\tau_i)^{-1}/\sum \tau_i^{-2}\right].\end{aligned}$$

Then, from (10), we get $R_0(\mathbf{1}) = 2\lambda\sum \tau_i(1+\lambda\tau_i)^{-3}S_0$, where

$$\begin{aligned}S_0 &= \frac{\sum 3\lambda\tau_i^2(1+\lambda\tau_i)^{-4}}{\sum \tau_i(1+\lambda\tau_i)^{-3}} - 2 + \frac{\sum 3\tau_i^{-2}(1+\lambda\tau_i)^{-1}}{\sum \tau_i^{-2}}\\ &= \frac{\sum \tau_i^2(2\lambda-\tau_i^{-1})(1+\lambda\tau_i)^{-4}}{\sum \tau_i(1+\lambda\tau_i)^{-3}} + \frac{\sum \tau_i^{-1}(2\tau_i^{-1}-\lambda)(1+\lambda\tau_i)^{-1}}{\sum \tau_i^{-2}}\\ &> \frac{\sum \tau_i^2(\lambda-\tau_i^{-1})(1+\lambda\tau_i)^{-4}}{\sum \tau_i(1+\lambda\tau_i)^{-3}} + \frac{\sum \tau_i^{-1}(\tau_i^{-1}-\lambda)(1+\lambda\tau_i)^{-1}}{\sum \tau_i^{-2}}. \qquad (33)\end{aligned}$$

Let $p_i = \tau_i(1+\lambda\tau_i)^{-3}$, $q_i = \tau_i^{-3}(1+\lambda\tau_i)^3$ and $r_i = (1-\lambda\tau_i)(1+\lambda\tau_i)^{-1}$. Then, clearly, $p_i > 0$, and the sequences $\{q_i\}$ and $\{r_i\}$ are non-constant and nonincreasing, so the discrete Chebyshev inequality (Mitrinović *et al.*, 1993, eq. (1.4), p. 240) gives $\sum p_i \sum p_i q_i r_i > \sum p_i q_i \sum p_i r_i$. Using this in (33), we obtain $S_0 > 0$ and hence $R_0(\mathbf{1}) > 0$, so $\mathbf{1}$ is not in RR.

Proof of lemma 2
The well-known expressions for $v(\lambda)$ and $b^2(\lambda)$ follow easily from their definitions. For the other expressions, we use the representation (Wahba, 1990, chap. 1) $f_\lambda = \sum d_j(\lambda)\theta_j + \sum c_i(\lambda)\xi_i$,

where $\{\theta_j\}$ is a basis for the space $\mathcal{H}_0$ of polynomials of degree $\leq m-1$ (i.e. the null space of d^m/dx^m) and $\{\xi_i\}$ is the set of representers of the evaluation functionals $\mathcal{H}_1 \to \mathbb{R}$, $f \to f(x_i)$. Here $\mathcal{H}_1 \subset \mathcal{W}^{m,2}[a,b]$ is the orthogonal complement of $\mathcal{H}_0$ with respect to the inner product $\sum_{k=0}^{m-1} f^{(k)}(a)g^{(k)}(a) + \int_a^b f^{(m)}(x)g^{(m)}(x)dx$. The vector of the coefficients $c_i(\lambda)$ is

$$\boldsymbol{c}(\lambda) = Q(Q^T(\Sigma + n\lambda I)Q)^{-1}Q^T\boldsymbol{y}, \tag{34}$$

where $\Sigma = [(\xi_i^{(m)}, \xi_j^{(m)})_{L_2}]$ and Q is an $n \times (n-m)$ matrix with orthogonal columns that are orthogonal to the vectors $(\theta_j(x_1), \ldots, \theta_j(x_n))^T$, $j = 1, \ldots, m$. The smoothing matrix can be written as $A(\lambda) = I - n\lambda Q(Q^T(\Sigma + n\lambda I)Q)^{-1}Q^T$. Using this expression and (34), it follows that

$$\|f_\lambda^{(m)}\|_{L_2}^2 = \boldsymbol{y}^T(n\lambda)^{-1}[I - A(\lambda) - (I - A(\lambda))^2]\boldsymbol{y} = (n\lambda)^{-1}\boldsymbol{y}^T[A(\lambda) - A^2(\lambda)]\boldsymbol{y},$$

and hence $v_1(\lambda) = \sigma^2(n\lambda)^{-1}\mathrm{tr}([A(\lambda) - A^2(\lambda)]) = \sigma^2\mu_{12}(\lambda)$.

From Kimeldorf & Wahba (1971), the interpolating spline f_{int} can be expressed in a form very similar to f_λ. In fact $f_{\mathrm{int}} = \sum \bar{d}_j\theta_j + \sum \bar{c}_i(0)\xi_i$, where $\bar{c}_i(0)$ is defined by (34) but with $\boldsymbol{y}$ replaced by $\boldsymbol{f}$. Therefore, using $E\boldsymbol{y} = \boldsymbol{f}$ and the expression for $A(\lambda)$, we obtain

$$\begin{aligned} b_1^2(\lambda) &= (E\boldsymbol{c}(\lambda) - \bar{\boldsymbol{c}}(0))^T\Sigma(E\boldsymbol{c}(\lambda) - \bar{\boldsymbol{c}}(0)) \\ &= n^{-1}\boldsymbol{f}^T(I - A(\lambda))^2 \lim_{\lambda\to 0}(\lambda^{-1}(I - A(\lambda)))\boldsymbol{f} \\ &= n^{-1}\sum (1 - a_{\lambda i})^2\tau_i(U^T\boldsymbol{f})_i^2 \end{aligned}$$

and the result follows.

Proof of lemma 3

If $-1/4 < r < 9/4$, then, from lemma 2 above and theorem 2.3 in Kou (2003), we have

$$\begin{aligned} b^2(\lambda) &= k\lambda^2 \sum (\tau_i^{-1} + \lambda)^{-2}\tau_i^{-r} \\ &= k\lambda^r \sum a_{\lambda i}^r (1 - a_{\lambda i})^{2-r} \\ &\sim k\lambda^r (1/(4\pi))B(r - 1/4, 9/4 - r)\lambda^{-1/4} \\ &= k\lambda^{r-1/4}(1/(4\pi))\Gamma(r - 1/4)\Gamma(9/4 - r), \end{aligned} \tag{35}$$

where B is the beta function $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$. (The parameter λ used in Kou (2003) is n times the parameter λ here.) Also, if $5/4 < r < 13/4$, then, from lemma 2 above and theorem 2.3 in Kou (2003), we have

$$\begin{aligned} b_1^2(\lambda) &= k\lambda^2 \sum (\tau_i^{-1} + \lambda)^{-2}\tau_i^{-r}\tau_i \\ &= k\lambda^{r-1} \sum a_{\lambda i}^{r-1} (1 - a_{\lambda i})^{3-r} \\ &\sim k\lambda^{r-1}(1/(4\pi))B(r - 5/4, 13/4 - r)\lambda^{-1/4} \\ &= k\lambda^{r-1-1/4}(1/(4\pi))\Gamma(r - 5/4)\Gamma(13/4 - r). \end{aligned} \tag{36}$$

Therefore assumption A4 holds with $p = r - 1/4$, where $5/4 < r < 9/4$.

Proof of theorem 3

From assumption A5, we have $EW(\hat{\lambda}_{\overline{V}}) \sim E\widetilde{W}(\hat{\lambda}_{\overline{V}})$ and $EW(\lambda_{EW}) \sim E\widetilde{W}(\lambda_{EW}) \geq E\widetilde{W}(\lambda_{E\widetilde{W}})$, where $\lambda_{E\widetilde{W}}$ minimizes $E\widetilde{W}(\lambda)$ for $\lambda \geq \alpha_n$. Therefore, from (26), we obtain

$$\frac{EW(\hat{\lambda}_{\overline{V}})}{EW(\lambda_{EW})} \leq \frac{E\widetilde{W}(\hat{\lambda}_{\overline{V}})}{E\widetilde{W}(\lambda_{E\widetilde{W}})}(1+o_P(1)) \leq \frac{E\|f_{\hat{\lambda}_{\overline{V}}} - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}}}{\min E\|f_\lambda - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}}}(1+o_P(1)), \tag{37}$$

and we can now estimate the right-hand side. For any $\lambda = \lambda(n)$ satisfying $\alpha_n \leq \lambda \to 0$ as $n \to \infty$, equation (27), lemma 2, assumption A4 and (15) yield

$$\begin{aligned} E\|f_\lambda - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}} &= b^2(\lambda) + v(\lambda) + b_1^2(\lambda) + v_1^2(\lambda) \\ &\sim c\lambda^p + \sigma^2\alpha l_2 n^{-1}\lambda^{-1/(2m)} + c_1\lambda^{p-1} + \sigma^2\alpha(l_1-l_2)n^{-1}\lambda^{-1-1/(2m)} \\ &\sim c_1\lambda^{p-1} + \sigma^2\alpha(l_1-l_2)n^{-1}\lambda^{-1-1/(2m)}. \end{aligned} \tag{38}$$

By minimizing this estimate and using the same argument as in lemma 3.2 of Nychka (1990), we get

$$\lambda_{E\widetilde{W}} = \left(\frac{\alpha(l_1-l_2)(2m+1)\sigma^2}{2mc_1(p-1)n}\right)^{2m/(2mp+1)}(1+o(1)). \tag{39}$$

Comparing (21) and (39), we define γ^* by the equation

$$\frac{\alpha l_2}{2mpc\gamma^*} = \frac{\alpha(l_1-l_2)(2m+1)}{2mc_1(p-1)}$$

giving (30). Using the estimates (38), (21) and (39) in the right-hand side of (37), and substituting $\gamma = \gamma^*\gamma/\gamma^*$ and simplifying, gives the formula (29) for K_1, and hence we have the bound (28).

If $\|f_{\text{int}} - f\|^2_{\widetilde{\mathcal{W}}} = o(\lambda^{p-1} + n^{-1}\lambda^{-1-1/(2m)})$ uniformly for all $\lambda \geq \alpha_n$, then, from (38) and (25), we have $E\widetilde{W}(\lambda) \sim E\|f_\lambda - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}}$, and it follows from assumption 5 that $EW(\lambda)$ behaves as in (38). By minimizing this estimate, we obtain the estimate for λ_{EW} in (31) (the same as for $\lambda_{E\widetilde{W}}$ in (39)). Then $EW(\lambda_{EW}) \sim EW(\lambda_{E\widetilde{W}}) \sim E\widetilde{W}(\lambda_{E\widetilde{W}})$. Using this and $E\widetilde{W}(\lambda) \sim E\|f_\lambda - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}}$ for $\lambda = \hat{\lambda}_{\overline{V}}$ and $\lambda = \lambda_{E\widetilde{\mathcal{W}}}$, it is clear that both of the inequalities in (37) can be replaced by equalities, giving $I_{EW}(\hat{\lambda}_{\overline{V}}) = K_1(1+o_P(1))$.

It is easy to see that if $\gamma^* \leq 1$, then $K_1 = K_1(\gamma)$, for $\gamma \in [0,1]$, has minimum value 1 at $\gamma = \gamma^*$. For class $\mathcal{C}$, from (35), (36) and (16), and using $\Gamma(z+1) = z\Gamma(z)$, we have $l_1/l_2 = 4/3$ and

$$\frac{c_1}{c} = \frac{\Gamma(r-5/4)\Gamma(13/4-r)}{\Gamma(r-1/4)\Gamma(9/4-r)} = \frac{9/4-r}{r-5/4} = \frac{2-p}{p-1}, \tag{40}$$

so the expression for γ^* in (30) simplifies to $\gamma^* = 0.6(2-p)/p$, which is in $(0, 0.6)$ for $p \in (1,2)$. For the weighted Sobolev risk $EW_\kappa(\lambda)$, by using the weighted inner products $(f,g)_{\widetilde{\mathcal{W}}_\kappa} = (\boldsymbol{f},\boldsymbol{g}) + \kappa(f^{(m)}, g^{(m)})_{L_2}$ and $(f,g)_{\mathcal{W}_\kappa} = (f,g)_{L_2(G)} + \kappa(f^{(m)}, g^{(m)})_{L_2}$, the argument above yields $E\|f_\lambda - f_{\text{int}}\|^2_{\widetilde{\mathcal{W}}_\kappa} \sim \kappa[c_1\lambda^{p-1} + \sigma^2\alpha(l_1-l_2)n^{-1}\lambda^{-1-1/(2m)}]$, and the inefficiency bound K_1 in (29) follows in the same way.