
http://researchrepository.murdoch.edu.au/8804/

Copyright: © 2008 Institute of Physics.

It is posted here for your personal use. No further distribution is permitted.
Strong Robust Generalized Cross-Validation for Choosing the Regularization Parameter

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

Abstract.
Let $f_\lambda$ be the Tikhonov regularized solution of a linear inverse or smoothing problem with discrete noisy data $y_i, i = 1, \ldots, n$. To choose $\lambda$ we propose a new strong robust GCV method denoted $R_1$GCV that is part of a family of such methods, including the basic RGCV method. $R_1$GCV chooses $\lambda$ to be the minimizer of $\gamma V(\lambda) + (1 - \gamma)F_1(\lambda)$, where $V(\lambda)$ is the GCV function, $F_1(\lambda)$ is a certain approximate total measure of the influence of each data point on $f_\lambda$ with respect to the regularizing norm or seminorm, and $\gamma \in (0, 1)$ is a robustness parameter. We show that $R_1$GCV is less likely to choose a very small value of $\lambda$ than both GCV and RGCV. RGCV and $R_1$GCV also have good asymptotic properties for general problems with independent errors. Strengthening previous results for RGCV, it is shown that the (shifted) RGCV and $R_1$GCV functions are consistent estimates of ‘robust risk’ functions, which place extra weight on the variance of $f_\lambda$. In addition RGCV is asymptotically equivalent to the modified GCV method. The results of numerical simulations for $R_1$GCV are consistent with the asymptotic results, and, for suitable values of $\gamma$, $R_1$GCV is more reliable and accurate than GCV.

*Subject Classifications:* AMS(2000) 65J20, 65D10, 62G08
1 Introduction

In many applications it is desired to estimate a function (or vector) \( f_0 \) from discrete noisy data \( y_i = L_i f_0 + \varepsilon_i, \ i = 1, \ldots, n \), where \( L_i \) are linear functionals and \( \varepsilon_i \) are random errors. For a linear operator equation \( Kf(x) = g(x) \) with data at \( x_i \) the functionals are \( L_i f = K f(x_i) \). A special case is the data smoothing problem, where \( L_i f = f(x_i) \). The general problem also includes a discretized operator equation or other finite dimensional linear model, in which case we have \( L_i f = (K f)_i \), where \( f \in \mathbb{R}^q, q \leq n \), and \( K \) is the \( n \times q \) model or design matrix.

To obtain a function estimate, we use Tikhonov regularization of the form [19]

\[
\text{minimize} \quad n^{-1} \sum_{i=1}^{n} (L_i f - y_i)^2 + \lambda \| Pf \|^2_W
\]  

over \( f \in W \), where \( W \) is an appropriate Hilbert space, e.g. a Sobolev space. The operator \( P : W \to W \) is either the identity or an orthogonal projection with finite dimensional null space. An important example is where \( \| Pf \|^2_W = \int (f^{(m)}(x))^2 \, dx \). In the case of a discrete linear model with \( L_i f = (K f)_i \) for \( f \in \mathbb{R}^q \), we replace \( \| Pf \|^2_W \) in (1.1) by \( \| M f \|^2 \), where \( \| \cdot \| \) is the Euclidean norm and, usually, either \( M \) is the identity or a finite difference operator of some order.

It is well known that the regularized solution \( f_{\lambda} \) of (1.1) depends critically on the choice of the regularization parameter \( \lambda \). A well-known method for choosing the parameter is generalized cross-validation (GCV) due to Wahba [18]. GCV is known to have favourable asymptotic properties and to work well in practice for large \( n \) (see [19, 10, 4] and the references there). However, for small or medium sized \( n \), the GCV method may not be reliable and can give a value of \( \lambda \) that is far too small (possibly even 0), corresponding to a very noisy regularized solution; see section 4.9 in [19] and [20, 3, 8].

Our aim is to develop methods that are both accurate and reliable for any sample size \( n \). In [13] we proposed the basic robust GCV method denoted RGCV (see also [15]). In this paper we define and investigate a whole class of robust GCV methods, which includes RGCV and a strong robust GCV method denoted R1GCV. We also show (Theorem 4.2) that a method based on a modified GCV function that has been used in [2, 6, 16] is asymptotically equivalent to the RGCV method.

All of these methods are defined in terms of the influence matrix \( A = A(\lambda) \) defined by \( Ay = L f_{\lambda} \), where \( L f = (L_1 f, \ldots, L_n f)^T \). The GCV choice of \( \lambda \) is the minimizer of the GCV function

\[
V(\lambda) = \frac{n^{-1} \| (I - A)y \|^2}{\mathbf{1}^T (I - A) \mathbf{1}} = \frac{n^{-1} \| (I - A)y \|^2}{(1 - \mu_1(\lambda))^2},
\]  

where \( \| \cdot \| \) is the Euclidean norm on \( \mathbb{R}^n \) and \( \mu_1(\lambda) = n^{-1} \text{tr} A \). The GCV function \( V(\lambda) \) is derived using a certain weighted sum of squared prediction errors \( y_k - L_k f_{\lambda}^{[k]} \), where \( f_{\lambda}^{[k]} \) is the regularized solution found by leaving out the \( k \)th data point.

The RGCV method in [13] is defined using the average influence \( n^{-1} \sum_k \| L f_{\lambda} - L f_{\lambda}^{[k]} \|^2 \), where \( \| L f_{\lambda} - L f_{\lambda}^{[k]} \|^2 \) is a measure of the influence of the \( k \)th data point on the regularized solution. Using a suitable approximation, the average influence becomes \( F(\lambda) = \mu_2(\lambda) V(\lambda) \), where \( \mu_2(\lambda) = n^{-1} \text{tr} (A^2) \). The RGCV estimate is defined to be the minimizer of the weighted sum \( \overline{V}(\lambda) = \gamma V(\lambda) + (1 - \gamma) F(\lambda) \), where \( \gamma \in (0, 1) \) is a robustness parameter.
The R₁ GCV method is defined by measuring the influence of the kth data point on $f_\lambda$ with a stronger norm as $\|P f_\lambda - P f_\lambda^{[k]}\|_{W}^2$. By summing these over $k = 1, \ldots, n$ and using a suitable approximation (see Section 3), we get the approximate total influence $F_1(\lambda) = \mu_{12}(\lambda) V(\lambda)$, where $\mu_{12}(\lambda)$ is defined as $\mu_{12}(\lambda) = (\mu_1(\lambda) - \mu_2(\lambda))/\lambda = -d\mu_1/d\lambda$. The R₁ GCV estimate is defined to be the minimizer of the weighted sum $\overline{V}_1(\lambda) = \gamma V(\lambda) + (1 - \gamma)F_1(\lambda)$, where $\gamma \in (0,1)$ is a robustness parameter. By considering the influence of the kth data point on $f_\lambda$ with respect to a certain class of seminorms $|| \cdot ||_\nu$, we obtain a whole family (R₁ GCV) of robust GCV methods, where R₂ GCV is equivalent to RGCV.

In Section 3 we show that the term $(1 - \gamma)F_1(\lambda)$ in $\overline{V}_1(\lambda)$ penalizes small values of $\lambda$, and $1 - \gamma$ determines the degree of this penalty. Theorem 3.1 shows that R₁ GCV is less likely to choose a very small value of $\lambda$ than either GCV or RGC, and hence it is a more robust method. Assuming the errors are independent with mean 0 and variance $\sigma^2$, the function $\mu_{12}(\lambda)$ in $\overline{V}_1(\lambda)$ is proportional to the variance $v_1(\lambda) \equiv E\|P f_\lambda - EP f_\lambda\|_{W}^2$. This means that R₁ GCV places much more weight on reducing the variability of the estimate $f_\lambda$ compared to GCV.

In Section 4 we show that RGC and R₁ GCV have favourable asymptotic properties as $n \to \infty$, which also help explain their robustness. We consider problems where $L_i f = K f(x_i)$, $x_i \in [0,1]$, for a linear operator $K$ (including $K = I$) and assume the errors $\xi_i$ are independent with mean 0 and common variance $\sigma^2$. Theorem 4.1 and Corollary 4.1 strengthen results in [13] and show that, in a neighbourhood of the optimal parameter, the RGC function $\overline{V}(\lambda)$ (with shift) is a consistent estimate of the ‘robust risk’, i.e. a weighted sum of the risk and its variance component. Theorem 4.3 and Corollary 4.2 show that, in a neighbourhood of the optimal parameter, the R₁ GCV function $\overline{V}_1(\lambda)$ (with shift) is a consistent estimate of the ‘strong robust risk’, i.e. a weighted sum of the risk and the variance $v_1(\lambda)$.

Section 5 describes numerical simulations with the R₁ GCV method for the discretized ill-posed problem of estimating the second derivative of a function $g(x)$ from noisy data $y_i = g(x_i) + \xi_i$, $i = 1, \ldots, n$. Even for moderately sized $n$ ($n = 51$) the numerical results are consistent with the results in Section 4. In contrast to GCV, the R₁ GCV method with $\gamma$ in a suitable range gives very accurate and reliable estimates of $\lambda$.

The under-smoothing problem of GCV depends strongly both on the smallness of $n$ and on the degree of correlation in the data [19]. It will be shown elsewhere that, unlike GCV, R₁ GCV performs well even when the data are correlated.

## 2 The regularized solution and loss functions

Assume that for each $i = 1, \ldots, n$, the linear functional $W \to \mathbb{R}$, $f \to L_i f$ is bounded and let $\eta_i$ be its representer, defined by $L_i f = (f, \eta_i)_W$ for all $f \in W$. Let $\{\theta_j, j = 1, \ldots, m\}$ be a basis for $N(P)$ and assume that $N(L) \cap N(P) = \{0\}$. It is well known [7, 19] that (1.1) has a unique solution which has the form

$$f_\lambda = \sum_{i=1}^{m} a_i \theta_i + \xi^T B^T (B \Sigma B^T + n \lambda I)^{-1} B y,$$

where $\xi = P \eta$, $\Sigma = [(\xi_i, \xi_j)_W]$ and $B$ is an $n \times n$ matrix with orthonormal rows whose span is $(LN(P))\perp$ (and $B = I$ if $P = I$). The influence matrix $A$, defined by $A y = L f_\lambda$, is given by

$$A = B \Sigma B^T + n \lambda I.$$
\[Ay = y - n\lambda B^T (BSB^T + n\lambda I)^{-1}By.\]

It is known [19] that the GCV function in (1.2) can be evaluated using the spectral decomposition \(n^{-1}BSB^T = \Lambda U^TU^T\), where \(U^TU = I\) and \(\Lambda = \text{diag}(\tilde{\lambda}_i)\) with \(\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \tilde{\lambda}_n \geq 0\). Define \(\tilde{\phi}_i = \sqrt{n} (B^TU)_i\) (ith column), \(i = 1, \ldots, n - m\), so \(n^{-1} (\tilde{\phi}_i, \tilde{\phi}_j) = \delta_{ij}\). Clearly \(A\) has eigenvalues \(\tilde{\lambda}_i/(\tilde{\lambda}_i + \lambda)\), with eigenvectors \(\tilde{\phi}_i\), \(i = 1, \ldots, n - m\), and also the eigenvalue 1 of multiplicity \(m\). Then the functions \(\mu_1(\lambda) = \text{tr} A\) and \(\mu_2(\lambda) = \text{tr}(A^2)\) can be expressed in terms of \(\tilde{\lambda}_i, i = 1, \ldots, n - m\) (see equations (2.9) and (2.10) in [13]). The function \(\mu_{12}(\lambda)\) in the definition of the R1 GCV method can also be expressed in terms of \(\tilde{\lambda}_i, i = 1, \ldots, n - m\), as

\[
\mu_{12}(\lambda) = (\mu_1(\lambda) - \mu_2(\lambda))/\lambda = -d\mu_1/d\lambda = n^{-1} \sum_{i=1}^{m} \tilde{\lambda}_i/(\tilde{\lambda}_i + \lambda)^2.
\] (2.2)

To measure the accuracy of a regularized solution, we will use the mean square prediction error \(R(\lambda) = n^{-1} \|Lf_x - Lf_0\|^2\) and a stronger loss function \(R_1(\lambda) = \|Lf_x - Lf_0\|^2\) that we now define. In fact we will define a whole class of loss functions. We assume here and throughout that \(\Sigma\) is nonsingular, so \(\tilde{\lambda}_i > 0\) for all \(i\).

In the case where \(P = I\), define for \(\nu \geq 0\) a class of inner products and norms on \(\mathbb{R}^n\) by \((w, z)_\nu = \sum_{i=1}^{n} 2^{-1} (w, \tilde{\phi}_i)(z, \tilde{\phi}_i)\tilde{\lambda}_i\nu\) and \(\|z\|_\nu^2 = (z, z)_\nu\). (These norms can be thought of as discrete versions of the Hilbert scale \(\{H_\nu\}\) function norms defined in [9, 10], where the \(H_0\) norm is an \(L_2\) norm.) Define the corresponding loss functions by \(R_\nu(\lambda) = \|Lf_x - Lf_0\|_\nu^2\). Note that for \(\nu = 0\), we have \(\|z\|_0^2 = n^{-1} \|z\|^2\), so \(\cdot\) \(\|\cdot\|_0\) is simply the normalized Euclidean norm and \(R_0(\lambda) = R(\lambda)\). The case with \(\nu = 1\) is also special as follows. Let \(P_\eta : W \rightarrow W\) be the orthogonal projection onto span\{\(\eta_\nu\}\}. Using the definition of \(\eta_\nu\), it is easy to show that \(P_\eta\) is given by \(P_\eta f = \eta^T \Sigma^{-1} Lf\). Then, since \(\Sigma^{-1} = n^{-1} U^T A^{-1} U^T\), we have

\[
\|P_\eta f\|_W^2 = (Lf)^T \Sigma^{-1} Lf = \sum_{i=1}^{n} 2^{-1} (Lf, \tilde{\phi}_i)^2 \tilde{\lambda}_i = \|Lf\|_1^2.
\] (2.3)

Using (2.3) with \(f = f_x - f_0\), and since \(P_\eta\) is orthogonal and \(P_\eta f_x = f_x\), we have

\[
\|Lf_x - Lf_0\|_W^2 = \|P_\eta f_x - P_\eta f_0\|_W^2 = \|f_x - f_0\|_W^2 - \|f_0 - P_\eta f_0\|_W^2.
\] (2.4)

Therefore \(R_1(\lambda) \equiv \|Lf_x - Lf_0\|_W^2\) is just a constant shift of \(\|f_x - f_0\|_W^2\) and has the same minimizer, so it is a very appropriate loss function. For example, if \(L_i f = K f(x_i)\) for an integral operator \(K\), it is known [17] that, under certain assumptions, the projection error \(\|f - P_\eta f\|_W = o(1)\) as \(n \rightarrow \infty\) with known rate. Then, from (2.3) and since \(P_\eta\) is orthogonal, we have \(\|Lf_x\|_W \sim \|f\|_W\) and, from (2.4), we obtain \(R_1(\lambda) \sim \|f_x - f_0\|_W^2\) as \(n \rightarrow \infty\).

In the case where \(P \neq I\), for any \(\nu \geq 0\) define a seminorm \(\|\cdot\|_\nu\) on \(\mathbb{R}^n\) by

\[
\|z\|_\nu^2 = \sum_{i=1}^{n} 2^{-1} (z, \tilde{\phi}_i)^2 \tilde{\lambda}_i \nu
\] (2.5)

and let \(R_\nu(\lambda) = \|Lf_x - Lf_0\|_\nu^2\). If \(\|z\|_\nu = 0\), then \((z, \tilde{\phi}_i) = \sqrt{n} U^T Bz = 0\) for all \(i\), so \(Bz = 0\) and hence \(z \in LN(P)\). Therefore \(\|\cdot\|_\nu\) is a norm on the orthogonal complement \((LN(P))^\perp\). With \(\nu = 0\) and \(z \in (LN(P))^\perp\), clearly \(\|z\|_0^2 = n^{-1} \|z\|^2\).

To examine the case with \(\nu = 1\), for any \(f \in W\), let \(f_{int}\) be the solution of the generalized interpolation problem: minimize \(\|Ph\|_W^2\) over \(h \in W\) subject to \(Lh = Lf\). As shown in [7], \(f_{int} = f_{\text{opt}}\) if \(f_{\text{opt}}\) is the solution of the problem \(\|f - Lf_0\|_W := 1\)\(\min_{f \in W}\{\|f - Lf_0\|_W : Lh = Lf\}\).
can be expressed in a form similar to $f_\lambda$ in (2.1), and $P f_{\text{int}} = \xi^T B^T (B \Sigma B^T)^{-1} BL f$. (Note that $f_{\text{int}} = P f$ when $P = I$.) Then $f \rightarrow P f_{\text{int}}$ is linear and, using $(B \Sigma B^T)^{-1} = n^{-1} U \Lambda^{-1} U^T$, we have

$$
\|P f_{\text{int}}\|_W^2 = (L f)^T B^T (B \Sigma B^T)^{-1} BL f = \sum_{i=1}^n n^{-2} (L f_i \tilde{\phi}_i)^2 \tilde{\lambda}_i^{-1} = \|L f\|_1^2. \tag{2.6}
$$

From the definition of $f_{\text{int}}$ it is clear that $(f_\lambda)_{\text{int}} = f_\lambda$, since minimizing $\|P h\|_W^2$ subject to $L h = L f_\lambda$ is equivalent to minimizing $n^{-1} \|L h - L f_\lambda\|^2 + \lambda \|P h\|_W^2$ subject to $L h = L f_\lambda$. Therefore, from (2.6) with $f = f_\lambda - f_0$, we have

$$
R_1(\lambda) \equiv \|L f_\lambda - L f_0\|_1^2 = \|P f_\lambda - P f_0\|_W^2. \tag{2.7}
$$

If $L f = K f(x_i)$ for an integral operator $K$, it is known [11] that, under certain assumptions, the error $\|f_0 - (f_0)_{\text{int}}\|_W \to 0$ as $n \to \infty$, so $R_1(\lambda)$ approximates $\|P f_\lambda - P f_0\|_W^2$.

In the discrete case we have the model $y = K f_0 + \varepsilon$ for $f_0 \in \mathbb{R}^q$, and wish to compare $f_\lambda$ with $f_0$. Clearly we can use the loss function $R(\lambda) = n^{-1} \|K f_\lambda - K f_0\|^2$. Also, using the SVD of $K$ if $M = I$ or the generalized SVD of $(K, M)$ if $M \neq I$, as in [13], we can define the loss functions $R_\nu(\lambda) = \|K f_\lambda - K f_0\|^2_\nu$ for $\nu \geq 0$. Here $\| \cdot \|^2_\nu$ is defined by an expression of the same form as (2.5), but with $n - m$ replaced by $n$ if $M = I$ and by $p$ if $M = M_{p \times q} \neq I$, $p \leq q \leq n$.

In the discrete case the functions $\mu_1(\lambda)$, $\mu_2(\lambda)$ and $\mu_{12}(\lambda)$ can be written in terms of eigenvalues $\tilde{\lambda}_i$, $i = 1, \ldots, p$ (see equations (2.13) and (2.14) in [13]).

### 3 Strong robust GCV method

Let $f_\lambda^{[k]}$ be the regularized solution of (1.1) obtained by leaving out the $k$th data point. As in [13], the ‘leaving-out-one’ lemma [19] gives

$$
y_k - L_k f_\lambda^{[k]} = (y_k - L_k f_\lambda) / (1 - a_{kk}) \quad \text{and} \quad L f_\lambda - L f_\lambda^{[k]} = A (y_k - L_k f_\lambda^{[k]}) e_k, \tag{3.1}
$$

where $e_k$ is the $k$th standard unit vector.

The basic robust GCV method RGCV was defined in [13] from (3.1) by taking the squared Euclidean norm $\|L f_\lambda - L f_\lambda^{[k]}\|^2$, which is a measure of the influence of the $k$th data point $(k, y_k)$ on the regularized solution. Using suitable approximations, it was found in [13] that the average of these influences is approximately $F(\lambda) \equiv \mu_2(\lambda) V(\lambda)$. The RGCV choice of $\lambda$ is the minimizer of

$$
\nabla F(\lambda) = \gamma V(\lambda) + (1 - \gamma) F(\lambda) = \frac{\gamma + (1 - \gamma) \mu_2}{(1 - \mu_1)^2} n^{-1} ||(I - A) y||^2 \tag{3.2}
$$

for some $\gamma \in (0, 1)$.

To derive the strong robust GCV method we measure the influence of $(k, y_k)$ in a stronger norm. For the regularized solution $f_\lambda$ of (1.1), define the influence of $(k, y_k)$ in the $W$ norm to be $\|P f_\lambda - P f_\lambda^{[k]}\|_W$. From (2.3) and since $P f_\lambda = f_\lambda$ when $P = I$, and from (2.6) and since $P f_\lambda^{[k]}_{\text{int}} = f_\lambda^{[k]}$ when $P \neq I$, we have $\|P f_\lambda - P f_\lambda^{[k]}\|_W = \|L f_\lambda - L f_\lambda^{[k]}\|_1^2$. Using (2.3) or (2.6), and (3.1), we get

$$
\|P f_\lambda - P f_\lambda^{[k]}\|_W^2 = \sum_{i=1}^n n^{-1} \sum_{k} (y_k - L_k f_\lambda)^2 \tilde{\lambda}_i^{-1} / (1 - a_{kk})^2 (\tilde{\lambda}_i + \lambda)^2.
$$
Define the total influence in the $W$ norm to be

$$
\sum_{k=1}^{n} \| Pf_k - P f^{[k]} \|_{W}^2 = \sum_{k=1}^{n} \sum_{i=1}^{m} n \frac{2(y_k - L_k f_k)^2 \tilde{\lambda}_i (\tilde{\phi}_i)_{k}^2}{(1 - a_{kk})^2 (\lambda_i + \lambda)^2}.
$$

(3.3)

Now, as in the derivation of the GCV function, we replace $a_{kk}$ in (3.3) by the average $n^{-1} \sum a_{kk} = n^{-1} \text{tr} A = \mu_1(\lambda)$. In addition we replace $\sum_{i=1}^{m} n^{-1} \tilde{\lambda}_i (\tilde{\phi}_i)_{k}^2 (\tilde{\lambda}_i + \lambda)^2$ in (3.3) by the average

$$
n^{-1} \sum_{i=1}^{m} \frac{\tilde{\lambda}_i (\tilde{\phi}_i)_{k}^2}{(\tilde{\lambda}_i + \lambda)^2} = n^{-1} \sum_{i=1}^{m} \frac{\tilde{\lambda}_i}{(\tilde{\lambda}_i + \lambda)^2} = \mu_{12}(\lambda),
$$

(3.4)

using (2.2), to get the approximate total influence in the $W$ norm

$$
F_1(\lambda) \equiv \mu_{12}(\lambda) n^{-1} \frac{\| (I - A)^{1/2} \|_{1}^2}{(1 - \mu_1(\lambda))^2} = \mu_{12}(\lambda) V(\lambda).
$$

(3.5)

For a discrete problem with regularized solution $f_k \in \mathbb{R}^n$, we define the influence of $(k, y_k)$ as $\| L f_k - L f_k^{[k]} \|_1$, where $\| \cdot \|_1$ is defined at the end of Section 2. Using the same steps as above, we can derive the same expression $F_1(\lambda)$ in (3.5) for the approximate total influence.

It is reasonable to expect that if $\lambda$ is too small, then because of the resulting sensitivity, the influences of some points $(k, y_k)$ will be large and so the total influence in (3.3) will be large. Therefore we use the approximate total influence $F_1(\lambda)$ as a penalty function. We define the strong robust GCV method (denoted R$_1$GCV) as: choose $\lambda$ to minimize the weighted sum

$$
\gamma V(\lambda) + (1 - \gamma) F_1(\lambda) = \frac{\gamma + (1 - \gamma) \mu_{12}(\lambda) n^{-1} \| (I - A)^{1/2} \|_{1}^2}{(1 - \mu_1(\lambda))^2}
$$

(3.6)

for some $\gamma \in (0, 1)$. Note that if $V(\lambda)$ is computed using a spectral decomposition, there is little extra work required to compute $\gamma V(\lambda)$ using the sum in (2.2) for $\mu_{12}(\lambda)$. Alternatively, from (2.2), no matter how the function $\mu_1(\lambda)$ is computed, $\mu_{12}(\lambda)$ can be estimated using a numerical derivative of $-\mu_1(\lambda)$.

When $\gamma = 1$ in (3.6) the method is just GCV. The parameter $\gamma$ is a robustness parameter; as $\gamma$ decreases the R$_1$GCV method becomes more robust, which can be seen as follows. As $\lambda \to \infty$, clearly $\mu_{12} \to 0$, so $(1/\gamma) \gamma V(\lambda) \sim V(\lambda)$. But as $\lambda \to 0$, we have

$$
(1/\gamma) \gamma V(\lambda) \sim [1 + ((1 - \gamma)/\gamma) \mu_{12}(0)] V(\lambda) \gg V(\lambda)
$$

if $\gamma$ is sufficiently small. In fact, since $\mu_{12}(0) = n^{-1} \sum \tilde{\lambda}_i^{-1}$ is usually very large, $(1/\gamma) \gamma V(\lambda) \gg V(\lambda)$ even for values of $\gamma$ that are just a bit smaller than 1. This means the term $(1 - \gamma) F_1(\lambda)$ in $\gamma V(\lambda)$ penalizes values of $\lambda$ that are close to 0, but not large values of $\lambda$. Also, since $\mu_{12}(0)$ is usually much larger than $\mu_2(0) = 1$, the function $F_1(\lambda)$ in R$_1$GCV is a stronger penalty function than $F(\lambda)$ in RGCV.

The following general result shows that for any $n$ and $\gamma \in (0, 1)$, R$_1$GCV is less likely than GCV and (usually) RGCV to choose a very small value of $\lambda$ (including 0).

**Theorem 3.1** Suppose that the errors $\varepsilon_i$ have a continuous probability distribution. For $\lambda \geq 0$, let $\overline{p}_i = P(\overline{\gamma}_i(\lambda) < 0)$, $\overline{p} = P(\overline{\gamma}_i(\lambda) < 0)$ and $p = P(\gamma(\lambda) < 0)$. For any $n$ and $\gamma \in (0, 1)$, we have $\overline{p}_i > p$ and $\overline{p} > p$. If $n^{-1} \sum \tilde{\lambda}_i^{-1} > 1$, then for all sufficiently small $\lambda \geq 0$, we also have $\overline{p}_i > \overline{p}$.
Proof. Using the definition (1.2) of $V(\lambda)$, it is not hard to see that for each $\lambda \geq 0$, the random variable $V'(\lambda)/V(\lambda)$ has a continuous distribution. From (3.2), $V'(\lambda) = z(\lambda)V(\lambda)$, where $z(\lambda) = \gamma + (1 - \gamma)\mu_2(\lambda)$, and let $a = z'(0)/z(0)$. Similarly, from (3.6), $V'(\lambda) = z(\lambda)V(\lambda)$, where $z_1(\lambda) = \gamma + (1 - \gamma)\mu_{12}(\lambda)$, and let $a_1 = z'_1(0)/z_1(0)$. Since $\mu_2(\lambda)$ and $\mu_{12}(\lambda)$ are strictly decreasing functions, clearly $a > 0$ and $a_1 > 0$ for all $\lambda \geq 0$. If $n^{-1}\sum_{i=1}^{n} \tilde{\lambda}_i^{-1} > 1$, then using this and the Cauchy-Schwarz inequality, we have $-z_1(0)z'(0) < 2(1 - \gamma)(n^{-1}\sum_{i=1}^{n} \tilde{\lambda}_i^{-1})^2 \leq -z(0)z'(0)$, so $a_1 > a$ for all sufficiently small $\lambda$. Then we obtain

$$
\overline{p}_1 = P(V'(\lambda)/V(\lambda) < a_1) > P(V'(\lambda)/V(\lambda) < a) = \overline{p} > P(V'(\lambda)/V(\lambda) < 0) = p. \quad \square
$$

In [13] it is shown that RGCV has the intuitive explanation that it places extra weight on the variance $v(\lambda) \equiv n^{-1}E\|L_{f,\lambda} - EL_{f,\lambda}\|^2$. The R1 GCV method has a similar interpretation. Assuming the errors are independent with mean 0 and variance $\sigma^2$, the function $\mu_{12}(\lambda)$ in $V_1(\lambda)$ in (3.6) is proportional to the variance $v_1(\lambda) \equiv E\|P_{f,\lambda} - EP_{f,\lambda}\|_W^2$ of $P_{f,\lambda}$ (see (4.5)), e.g., $v_1(\lambda) = \|f'^{\lambda}_1 - EF^{f,\lambda}\|_2^2$, where $\cdot$ $\|_2$ is the $L_2$ norm. Since this variance is measured in the strong $W$ norm, R1 GCV places much more weight on reducing the variability of $f_\lambda$ compared to GCV (when $\gamma = 1$). Asymptotic results in Section 4.2 confirm this behaviour.

Geometrically, it is clear that the term $(1 - \gamma)F_1(\lambda)$ in (3.6) has the effect of modifying the shape of $V(\lambda)$ for very small $\lambda$ so that $V_1(\lambda)$ better defines a suitable global minimizer (compare Figure 3 in [13] and Figure 2 in Section 5).

An important practical question is how one should select the robustness parameter $\gamma$. Here we only mention a simple approach of taking the largest value of $\gamma$ for which the (modified) graph of $V_1(\lambda)$ defines a ‘clear’ positive global minimizer. For each value of $\gamma$ in a decreasing sequence starting at 1, we compute the minimizer of $V_1(\lambda)$ and monitor for abrupt change to the minimizer (as discussed for RGCV in [13]). We select $\gamma$ as the largest value in the sequence below which there are no abrupt changes and for which the function $V_1(\lambda)$ (or log $V_1(\lambda)$) satisfies some specified minimum threshold for the curvature at its minimum. It would also be useful to plot the sequence of functions $V_1(\lambda)$ to check their behaviour by eye.

Note that since usually $\mu_{12}(\lambda)$ is much larger than $\mu_2(\lambda)$ for small $\lambda$, the value of $\gamma$ for R1 GCV need not be anywhere near as small as that for RGCV to achieve the same effect of penalizing small $\lambda$ values. In fact, usually the value of $\gamma$ in R1 GCV should not be taken too far from 1 because otherwise the method may choose $\lambda$ too large, leading to an inaccurate regularized solution. We will quantify this at the end of Section 4 using some asymptotic results, and it will also be seen in the simulations of Section 5.

By considering the influence of $(k,y_k)$ on $f_\lambda$ with respect to the squared norm or seminorm $\|\cdot\|_p^2$ in (2.5), we have

$$
\|L_{f,\lambda} - L_{f,\lambda}^{[k]}\|_p^2 = \sum_{i=1}^{n} \frac{n}{n} (L_{f,\lambda} - L_{f,\lambda}^{[k]})(\tilde{\phi})_i^2 \tilde{\lambda}_i^\nu,
$$

and following the same approach as above, we can derive a whole family of robust GCV methods. The resulting method, denoted $R_\nu$ GCV, chooses $\lambda$ to minimize $V_\nu(\lambda) = \gamma V(\lambda) + (1 - \gamma)F_\nu(\lambda)$, where $F_\nu(\lambda) = \mu_{12}(\nu,\gamma,2)(\lambda)V(\lambda)$ and $\mu_{12}(\nu,\gamma,2)(\lambda) = n^{-1}\sum_{i=1}^{n} \tilde{\lambda}_i^\nu (\tilde{\lambda}_i + \lambda)^2$. With $\nu = 1$, this is the R1 GCV method above. With $\nu = 0$, the function $\mu_{12}(\nu,\gamma,2)(\lambda)$ differs from $\mu_2(\lambda)$ only by the
constant $n^{-1}m$, so the corresponding method is equivalent to the RGCV method (with a slightly different parameter $\gamma$). If, as is usually the case, $\mu(\nu) = n^{-1}\sum \tilde{\lambda}_i \nu$ increases with $\nu$, then, using the same reasoning as for $R_{1GCV}$ above, the larger the value of $\nu$, the more robust is the $R_pGCV$ method.

Focusing on $R_{1GCV}$ and $R_{1GCV}$, a natural question is which method should be used in practice. Since both methods are not much more computationally expensive than GCV, it is recommended to first use RGCV, and then use $R_{1GCV}$ if it appears (from the curvature of the RGCV function) that more stabilization is needed.

4 Asymptotic properties of robust GCV methods

4.1 Asymptotic behaviour of RGCV

Let $R(\lambda)$ be the mean square prediction error, and so $ER(\lambda)$ is the risk. Define the robustified loss function $\overline{R}(\lambda)$ as $\overline{R}(\lambda) = \gamma R(\lambda) + (1-\gamma)v(\lambda)$, where $v(\lambda) = n^{-1}E\|L\hat{f}_\lambda - E L\hat{f}_\lambda\|^2$ is the variance.

Also define the ‘robust risk’ as $E\overline{R}(\lambda)$. If the errors are independent with mean 0 and variance $\sigma^2$, then $v(\lambda) = \sigma^2 \mu_2(\lambda)$ and the risk is $ER(\lambda) = \hat{b}^2(\lambda) + v(\lambda)$, where $\hat{b}^2(\lambda) = n^{-1}\|E L\hat{f}_\lambda - L f_0\|^2$ is the squared bias. In [13] it is shown that the expected RGCV function $E\overline{R}(\lambda)$ behaves asymptotically like $E\overline{R}(\lambda) - \gamma\hat{b}^2$. Here we strengthen the analysis by considering the RGCV function $\overline{R}(\lambda)$ itself.

Denote $\hat{\sigma}^2 = n^{-1}e^T e$. We will show that $\overline{R}(\lambda) - \gamma\hat{\sigma}^2$ is a consistent estimate of $\overline{R}(\lambda)$. We use a result of Gu, Theorem 3.3 in [4], which shows the consistency of the GCV function in a simple way. The result is given in [4] for spline smoothing but it is easy to see that it also applies in the general setting here.

Condition C1. $n E R(\lambda) \to \infty$ as $n \to \infty$.
Condition C2. $\mu_2^2(\lambda)/\mu_2(\lambda) \to 0$ as $n \to \infty$.
Condition C3. $\mu_2(\lambda) \to 0$ as $n \to \infty$.

Theorem 4.1 Assume the errors $\varepsilon_i$ are independent with mean 0 and variance $\sigma^2$, and have uniformly bounded fourth moments. If Conditions C1, C2 and C3 hold, then as $n \to \infty$,

$$\overline{R}(\lambda) - \gamma\hat{\sigma}^2 = o_p(\overline{R}(\lambda)) \quad \text{and} \quad \overline{R}(\lambda) - E\overline{R}(\lambda) = o_p(E\overline{R}(\lambda)).$$

Proof. Using the definitions of $\overline{R}(\lambda)$ and $\overline{R}(\lambda)$, and rearranging, we have

$$\overline{R}(\lambda) - \gamma\hat{\sigma}^2 = \gamma (V(\lambda) - R(\lambda) - \hat{\sigma}^2) + (1-\gamma)\mu_2(\hat{\sigma}^2 - \sigma^2) + (1-\gamma)\mu_2 R(\lambda). \quad (4.1)$$

From Theorem 3.3 in [4], $V(\lambda) - R(\lambda) - \hat{\sigma}^2 = o_p(R(\lambda))$ and so, since $\mu_2 \leq 1$, the first term on the right-hand side in (4.1) is $o_p(R(\lambda))$. As in the proof of Theorem 3.1 in [4], we have $R(\lambda) - ER(\lambda) = o_p(R(\lambda))$, and hence, since $\mu_2 \leq (1/\sigma^2)ER(\lambda)$ and $\hat{\sigma}^2 - \sigma^2 = O_p(n^{-1/2})$, the second term is also $o_p(R(\lambda))$. Since $\mu_2 \to 0$, the last term in (4.1) is $o_p(R(\lambda))$. The first result follows since $R(\lambda) \leq (1/\gamma)\overline{R}(\lambda)$. The second result follows directly from $R(\lambda) - ER(\lambda) = o_p(ER(\lambda))$. \hfill \Box

Corollary 4.1 shows that Conditions C1, C2 and C3 hold uniformly in a neighbourhood of the optimal regularization parameter in the general framework of [10, 13], where $L_i f = K f(x_i)$, $x_i \in [0,1]$, and $K$ is a linear operator (including $K = I$). The errors $\varepsilon_i$ are assumed to be
independent with mean 0 and variance $\sigma^2$. The assumptions made on the points $x_i$ and on the operator $K : W \to L^2$ are given in Assumptions 4.2-4.5 in [13] and for brevity are not repeated here. In particular it is assumed that the eigenvalues $\tau_i$, defined by $P \psi_i = \tau_i K^* K \psi_i$, satisfy $c_1 \tau_i^r \leq \tau_i \leq c_2 \tau_i^r$ for all $i$, with constants $c_1, c_2 > 0$.

Under these assumptions there is a sequence $\alpha_n \to 0$ such that for $\lambda \geq \alpha_n$ the squared bias $b^2(\lambda)$ and variance $v(\lambda)$ components of the risk $ER(\lambda)$ can be estimated. The asymptotic behaviour of $b^2(\lambda)$ depends on the smoothness of $f_0$ measured in a class of norms $\| \cdot \|_{\beta}$ on spaces $W_\beta$ defined in [10, 13]. From Proposition 4.1 in [13], if $f_0 \in W_\beta$, the minimum over $\lambda \geq \alpha_n$ of the resulting upper bound on $ER(\lambda)$ occurs at $\lambda \approx \lambda_R^*$ as $n \to \infty$, where

$$\lambda_R^* = \begin{cases} (\sigma^2 n^{-1/r})^{(\beta + 1)/2}, & 1 \leq \beta < 2, \\ (\sigma^2 n^{-1/r})^{(\beta + 1)/2}, & \beta \geq 2, \end{cases}$$

(4.2)

assuming that $\lambda_R^* \geq \alpha_n$. (We use the notation $a_n \approx b_n$ to mean $c_1 b_n \leq a_n \leq c_2 b_n$ for some positive constants $c_1$ and $c_2$.) If $f_0 \in W_\beta$, $\beta \geq 2$, and $P f_0 \neq 0$, then the minimizer $\lambda_R = \lambda_R(n)$ of $ER(\lambda)$ over $\lambda \geq \alpha_n$ satisfies $\lambda_R \approx \lambda_R^*$ and $ER(\lambda_R) \approx ER(\lambda_R^*)$. Because the robust risk $ER(\lambda) = \gamma b^2(\lambda) + v(\lambda)$ is just a different weighted sum of $b^2(\lambda)$ and $v(\lambda)$, the minimizer of $ER(\lambda)$ has the same asymptotic behaviour as the minimizer of $ER(\lambda)$.

**Corollary 4.1** Suppose that Assumptions 4.2-4.5 in [13] hold and $\alpha_n \to 0$ is the sequence defined in Proposition 4.1 in [13]. Assume that $P f_0 \neq 0$ and the errors $\varepsilon_i$ are independent with mean 0 and variance $\sigma^2$, and have uniformly bounded fourth moments. Then Conditions C1, C2 and C3 hold uniformly for all $\lambda \geq \alpha_n$ satisfying $n^{1/\lambda} \to 0$. Moreover, the conclusions of Theorem 4.1 hold in a neighbourhood of $\lambda_R^*$.

**Proof.** From Theorems (4.1) and (4.3) in [10], as $n \to \infty$,

$$\mu_1(\lambda) \approx n^{1/r} D(\lambda; -1/r, -1) \quad \text{and} \quad \mu_2(\lambda) \approx n^{1/r} D(\lambda; -1/r, -2)$$

(4.3)

uniformly in $\lambda \in [\alpha_n, \infty)$, where $D(\lambda; a, b) \equiv \lambda^a$, if $\lambda \leq 1$, and $D(\lambda; a, b) \equiv \lambda^b$, if $\lambda > 1$. Combining the resulting estimate of $v(\lambda) = \sigma^2 \mu_2(\lambda)$ with the lower bound $b^2(\lambda) \geq \gamma^2$ in (4.4) in [13] (since $P f_0 \neq 0$), it follows easily that Condition C1 holds uniformly for $\lambda \geq \alpha_n$. If also $n^{1/\lambda} \to 0$, then from (4.3), Conditions C2 and C3 hold uniformly. The last part follows by using (4.2) in $n^{1/\lambda} \to 0$. $\square$

Another method that has been used to correct the under-smoothing problem of GCV is the modified GCV method [2, 6, 16], which selects $\lambda$ as the minimizer of

$$V(\lambda) = n^{1/r} \frac{\| (I - A) y \|^2}{\| (I - c A) y \|^2} = n^{1/r} \frac{\| (1 - A) y \|^2}{\| (1 - c \mu_1(\lambda)) y \|^2},$$

where $c > 1$ is a stabilization parameter. The factor $c$ has the effect of limiting the degrees of freedom in the regularized solution, with a larger value giving greater stabilization.

From Theorem A.1 in [2] it is clear that if $\mu_1(\lambda)/\mu_2(\lambda)$ approaches a constant as $n \to \infty$, then $EV(\lambda)$ behaves asymptotically like $(1/\gamma) E \overline{\lambda}^r(\lambda)$ for a certain value of $\gamma$. In the next result we strengthen this to show that the modified GCV method is asymptotically equivalent to the RCGCV method.
Theorem 4.2 Assume the errors $\varepsilon_i$ are independent with mean 0 and variance $\sigma^2$, and have uniformly bounded fourth moments. Suppose that Condition C1 holds and also $\mu_1(\lambda) \to 0$ and $\mu_1(\lambda)/\mu_2(\lambda) \to M$, a constant, as $n \to \infty$ (so C2 and C3 hold). If $1/\gamma = 1 + 2(c - 1)M$, then

$$(1/\gamma)\nabla(\lambda) - V_\varepsilon(\lambda) = o_P(\mathcal{R}(\lambda)) \quad \text{and} \quad V_\varepsilon(\lambda) - (1/\gamma)\overline{\mathcal{R}}(\lambda) - \hat{\sigma}^2 = o_P(\mathcal{R}(\lambda)).$$

Proof. Let $t = (1 - \gamma)/\gamma$. Since $\mu_1(\lambda) \to 0$, binomial expansions yield

$$(1/\gamma)\nabla(\lambda) - V_\varepsilon(\lambda) = n^{-1}|| (I - A) y ||^2 [t \mu_2 + 2t^2 \mu_1 \mu_2 + 2t \mu_1 + 2c \mu_1 + O(\mu_1^2)].$$

Substituting $t = 2(c - 1)M$, the right-hand side is $n^{-1}|| (I - A) y ||^2 O(\mu_1^2)$. After a simple rearrangement, Theorem 3.1 in [4] gives

$$n^{-1}|| (I - A) y ||^2 O(\mu_1^2) = [o_P(R(\lambda)) + R(\lambda) + \sigma^2(1 - 2\mu_1) + (\hat{\sigma}^2 - \sigma^2)]O(\mu_1^2).$$

(4.4)

As in the proof of Theorem 4.1, we have $\hat{\sigma}^2 - \sigma^2 = O_P(n^{-1/2})$ and $\mu_1 \sim M^2 \mu_2 = o_P(R(\lambda))$, so the first result follows since $R(\lambda) \leq (1/\gamma)\overline{\mathcal{R}}(\lambda)$. The second result follows easily from the first result and Theorem 4.1. \(\Box\)

In the framework of [13], if the eigenvalues $\tau_i$ satisfy $\tau_i = kt^r$ for some constant $k$, then it is known (see p. 52 of [10] and p. 167 of [12]) that $\mu_1(\lambda)/\mu_2(\lambda) \to M = r/(r - 1)$ as $n \to \infty$. In the case of cubic spline smoothing it is known that $\tau_i = kt^r$ with $r = 4$, so $M = 4/3$ and the relation between $\gamma$ and $c$ is $1/\gamma = 1 + 8(c - 1)/3$. Empirical results for spline smoothing in [2, 6] suggest that a good value of $c$ is in the range $[1.2, 1.4]$, which corresponds to $\gamma$ in the range $[0.48, 0.65]$. Our experiments with RGCV (e.g. in [13]) suggest that good results can be obtained for $\gamma$ in a larger interval, say $[0.1, 0.6]$.

4.2 Asymptotic behaviour of $R_1$ GCV

Consider the variance $v_1(\lambda)$ of $P f_\lambda$ in the $W$ norm. Under the assumption that the errors $\varepsilon_i$ are independent with mean 0 and common variance $\sigma^2$, $v_1(\lambda)$ satisfies (see A.7 and A.8 in [10])

$$v_1(\lambda) \equiv E\|P f_\lambda - EP f_\lambda\|_W^2 = \sigma^2 \mu_1(\lambda) .$$

(4.5)

Define the strong robustified loss function $\overline{R}_1(\lambda)$ as $\overline{R}_1(\lambda) = \gamma R(\lambda) + (1 - \gamma)v_1(\lambda)$. Define the ‘strong robust risk’ to be

$$E\overline{R}_1(\lambda) = \gamma ER(\lambda) + (1 - \gamma)v_1(\lambda) = \gamma b^2(\lambda) + \gamma v(\lambda) + (1 - \gamma)v_1(\lambda).$$

(4.6)

Clearly $E\overline{R}_1(\lambda)$ is a modification of $ER(\lambda)$ that puts more weight on controlling the variance, and, since $v_1(\lambda)$ is measured with a stronger norm than $v(\lambda)$, $E\overline{R}_1(\lambda)$ is stronger than $ER(\lambda)$ in this respect. An example is $v_1(\lambda) = E\|f_\lambda^{(n)} - f_\lambda^*\|_2^2$, where $\| \cdot \|_2$ is the $L_2$ norm.

Note that the strong robust risk $E\overline{R}_1(\lambda)$ is a reasonable error function since, assuming $f_\lambda \in W$, it behaves like $\gamma ER(\lambda) + (1 - \gamma)E\|P f_\lambda - P f_\lambda\|_W^2$ for small $\lambda$ (where the variance $v_1(\lambda)$ is large and the squared bias $\|EP f_\lambda - P f_\lambda\|_W^2$ is small), and like $\gamma ER(\lambda)$ for large $\lambda$ (where $v_1(\lambda)$ is small).

The next result shows that $\nabla(\lambda) - \gamma \hat{\sigma}^2$ is a consistent estimate of $\overline{R}_1(\lambda)$, where $\hat{\sigma}^2 = n^{-1}\varepsilon^T \varepsilon$.

The proof is the same as that of Theorem 4.1.

Condition C4. $\mu_{12}(\lambda) \to 0$ as $n \to \infty$. 

9
Theorem 4.3 Assume the errors $\varepsilon_i$ are independent with mean 0 and variance $\sigma^2$, and have uniformly bounded fourth moments. If Conditions C1, C2 and C4 hold, then as $n \to \infty$,

$$\nabla_1(\lambda) - \nabla_1(\lambda) - \gamma\hat{\sigma}^2 = o_p(\nabla_1(\lambda)) \quad \text{and} \quad \nabla_1(\lambda) - E\nabla_1(\lambda) = o_p(E\nabla_1(\lambda)).$$

Clearly there is a similar result for any $R_\nu$ GCV method; under Conditions C1, C2 and the condition $\mu(\nu, \nu, \nu(\lambda)) \to 0$, the function $\nabla_\nu(\lambda) - \gamma\hat{\sigma}^2$ is a consistent estimate of $\nabla_\nu(\lambda)$, where $\nabla_\nu(\lambda) = \gamma R(\lambda) + (1 - \gamma)\nu(\lambda)$ and $\nu(\lambda) \equiv E\|f_\lambda - E\|_{\nu}$ with $\| \cdot \|_\nu$ defined in (2.5).

Corollary 4.2 shows that in the framework of [10, 13], the Conditions C1, C2 and C4 hold in a neighbourhood of the optimal regularization parameter. First we state a known asymptotic result for the $W$ norm error function $E\|f_\lambda - f_0\|_W^2$. Under the assumption of independent errors with mean 0 and variance $\sigma^2$, this error function can be decomposed into the squared bias $\|E f_\lambda - f_0\|_W^2$ and variance $E\|f_\lambda - E f_\lambda\|_W^2$. Under Assumptions 4.1, 4.2, 4.3 and 4.5 in [13], these terms can be estimated [1, 9, 14]. From Corollary 3.1 in [10] with $\rho = 1$, if $f_0 \in W_\beta$, $\beta > 1$, the minimum of the resulting upper bound on $E\|f_\lambda - f_0\|_W^2$ occurs at $\lambda \approx \lambda^*_W$, where

$$\lambda^*_W = \begin{cases} (\sigma^2 n^{1/2})^{r/(\beta r + 1)}, & 1 < \beta < 3, \\ (\sigma^2 n^{1/2})^{r/(\beta r + 1)}, & \beta \geq 3, \end{cases}$$

assuming that $\lambda^*_W \geq \alpha_n$ for a certain sequence $\alpha_n \to 0$. Also, if $f_0 \in W_\beta$, $\beta \geq 3$, then the minimizer $\lambda_W$ of $E\|f_\lambda - f_0\|_W^2$ over $\lambda \geq \alpha_n$ satisfies $\lambda_W \approx \lambda^*_W$. Note that $\lambda^*_W \geq \lambda^*_R$, where $\lambda^*_R$ is defined in (4.2), and $\lambda^*_W > \lambda^*_R$ if $\beta > 2$.

If the function $f_0$ is sufficiently ‘smooth’, the minimizer of $E\nabla_1(\lambda)$ in (4.6) will satisfy $\lambda \approx \lambda^*_W$ as $n \to \infty$. This follows from the results Theorem 4.5 and Proposition 3.1 in [10] which show that, if $f_0 \in W_\beta$, $\beta \geq 3$, then both $b^2(\lambda) \approx \min\{1, \lambda^2\}$ and $\|E f_\lambda - f_0\|_W^2 \approx \min\{1, \lambda^2\}$, uniformly for $\lambda \geq \alpha_n$. Also, from (4.3) and (4.8), if $\lambda = o(1)$ then clearly $\gamma \mu_2(\lambda) + (1 - \gamma)\mu_1(\lambda) \approx \mu_1(\lambda)$, uniformly for $\lambda \geq \alpha_n$. If $f_0$ is less ‘smooth’, meaning that $f_0 \in W_\beta$ for $2 \leq \beta < 3$, then from Theorem 4.5 and Proposition 3.1 in [10], it is likely that the minimizer of $E\nabla_1(\lambda)$ will decay faster to $0$, somewhere between $\lambda^*_W$ and $\lambda^*_R$.

Corollary 4.2 Under the same assumptions as in Corollary 4.1, the Conditions C1, C2 and C4 hold uniformly for all $\lambda \geq \alpha_n$, satisfying $n \lambda^{1 + 1/r} \to 0$. Moreover, if $f_0 \in W_\beta$, $\beta > 1$, the conclusions of Theorem 4.3 hold in an interval containing $\lambda^*_W$ and $\lambda^*_R$.

Proof. Conditions C1 and C2 hold uniformly as in Corollary 4.1. Condition C4 holds uniformly from the estimate (see Theorem 4.2 in [10])

$$\mu_1(\lambda) \approx n^{1/2} D(\lambda; 1 + 1/r, -2),$$

where $D$ is defined after (4.3). The last part follows by using (4.2) and (4.7) in $n \lambda^{1 + 1/r}$.

Since the $R_1$ GCV function $\nabla_1(\lambda) = \gamma V(\lambda) + (1 - \gamma)F_1(\lambda)$ involves $F_1(\lambda) = \mu_1(\lambda) V(\lambda)$ as a penalty function, it is of interest to know how $F_1(\lambda)$ behaves. The next result shows that $F_1(\lambda)$ tracks the strong error function $E\|P f_\lambda - P f_0\|_W^2$ for a set of $\lambda$ values that are asymptotically a bit smaller than $\lambda^*_W$ in (4.7), but which includes $\lambda^*_R$ in (4.2) if $f_0$ is sufficiently ‘smooth’.
Theorem 4.4  Suppose the assumptions in Corollary 4.1 hold and \( f_0 \in W_\beta, \beta \geq 1 \). If \( \lambda = \lambda(n) \to 0 \) as \( n \to \infty \) such that \( \lambda \geq \alpha_n, n\lambda^{1/r} \to \infty \) and either \( \lambda^2(n\lambda^{1/r}) \to 0 \) if \( 1 \leq \beta < 3 \) or \( \lambda^3(n\lambda^{1/r}) \to 0 \) if \( \beta \geq 3 \), then

\[
F_1(\lambda) = E\|P f_\lambda - P f_0\|_W^2 (1 + o_P(1)).
\]

In particular, the result holds for \( \lambda = \lambda_R \) in (4.2) if \( \beta > 2 \) (assuming \( \lambda_R \geq \alpha_n \)), since \( \lambda_R^2(n\lambda_R^{1/r}) \approx n \beta^{(2r)/(\beta r+1)} \to 0 \) and \( \lambda_R^3(n\lambda_R^{1/r}) \to 0 \). The result also holds for \( \lambda = (\lambda_W) \), \( 1 < t < t = \min(\beta, 3) + 1/r \), (assuming \( \lambda \geq \alpha_n \)) but not for \( \lambda = \lambda_W \) in (4.7), since \( \lambda_W^2(n\lambda_W^{1/r}) = \sigma^2 \) if \( 1 < \beta < 3 \), and \( \lambda_W^3(n\lambda_W^{1/r}) = \sigma^2 \) if \( \beta \geq 3 \).

Proof. Decomposing \( E\|P f_\lambda - P f_0\|_W^2 \) into the squared bias \( b_1^2(\lambda) = \|EP f_\lambda - P f_0\|_W^2 \) and variance \( v_1(\lambda) = \sigma^2 \mu_1(\lambda) \), we can write

\[
F_1(\lambda) = E\|P f_\lambda - P f_0\|_W^2 = \mu_1(\lambda) - v_1(\lambda) - b_1^2(\lambda).
\]  

(4.9)

As in the proof of Theorem 4.1, we have \( V(\lambda) - \sigma^2 = O_P(V(\lambda) + n^{1/2}) \). Substituting this and \( V(\lambda) = b_1^2 + \sigma^2 \mu_2 \) into (4.9), and using \( E\|P f_\lambda - P f_0\|_W^2 \geq \sigma^2 \mu_1 \), we get

\[
F_1(\lambda) - E\|P f_\lambda - P f_0\|_W^2 = (b_1^2 + \sigma^2 \mu_2 + n^{1/2} + b_1^2/\mu_1)\sigma^2 E\|P f_\lambda - P f_0\|_W^2 O_P(1).
\]  

(4.10)

From Proposition 3.1 in [10] with \( \rho = 1 \), the squared bias \( b_1^2(\lambda) \) can be bounded as

\[
b_1^2(\lambda) \leq \|E f_\lambda - f_0\|_W^2 \leq \left\{
\begin{array}{ll}
c_1 \min\{1, \lambda^{1/r}\} f_0\|_W^2, & 1 \leq \beta < 3,
\end{array}
\right.
\]  

(4.11)

uniformly for \( \lambda \geq \alpha_n \). Substituting (4.3), (4.8), (4.11) and the bound on \( b_1^2(\lambda) \) in [13] all into (4.10), and using the conditions on \( \lambda = \lambda(n) \), it is not hard to show that the right-hand side of (4.10) is \( E\|P f_\lambda - P f_0\|_W^2 O_P(1) \) as \( n \to \infty \). The result follows. \( \square \)

Lastly in this section we investigate the choice of the robustness parameter \( \gamma = \gamma_1 \) in RGCV by relating it to the robustness parameter \( \gamma = \gamma_0 \) in RCGCV. In the framework of [10, 13], from Corollary 4.2, we have \( \gamma_1(\lambda) - \gamma_0^2 = E\gamma_1(\lambda)(1 + o_P(1)) \), where \( E\gamma_1(\lambda) \) is given in (4.6). If \( f_0 \in W_\beta, \beta \geq 2 \), then \( b_1^2(\lambda) \approx \lambda^2 \), and here we assume further that \( b_1^2(\lambda) = a\lambda^2 \) for some constant \( a \). Also assume here that the eigenvalues \( \tau_i \) satisfy \( \tau_i = kv_i \) for some constant \( k \).

Then it is known (see p. 52 of [10] and p. 167 of [12]) that \( v(\lambda) \sim c k^{-1/r} \sigma^2 n^{1/\beta} \), where \( c = \int_0^\infty (1 + x^r) dx = (1/r)(1 - 1/r) \Gamma(1/\beta r)(1 - 1/r)/\Gamma(2) \), and \( v(\lambda) \sim d k^{-1/r} \sigma^2 n^{1/\beta} \). By simple calculus the minimizer of the resulting estimate of \( V_1(\lambda) \) with \( \gamma = \gamma_1 \) satisfies

\[
2ar\gamma_1^{1/r}(\beta + d(r + 1)(1 - \gamma_1)) = \sigma^2 n^{1/\beta}.
\]

(4.12)

For RGCVC, from Corollary 4.1 we have \( V(\lambda) - \gamma_0^2 = E\gamma_0(\lambda)(1 + o_P(1)) \). With the same assumptions on \( b_1^2(\lambda) \) and \( \tau_i \) as above, the minimizer of the resulting estimate of \( V(\lambda) \) with \( \gamma = \gamma_0 \) satisfies \( 2ar\gamma_0^{1/r}/c = \sigma^2 n^{1/\beta} \). By equating this to (4.12) we find that, if \( R_1 \) with \( \gamma = \gamma_1 \) is to choose the same ‘near optimal’ parameter \( \lambda \) as \( R_1 \) with \( \gamma = \gamma_0 \), then

\[
\gamma_1 = \gamma_0/[\gamma_0 (1 - \gamma_0)(r - 1)/(r + 1)].
\]

(4.13)
From (4.13) it follows that, if \( \lambda \ll \gamma_0 \) as we can expect, then \( \gamma_1 \) will be significantly larger than \( \gamma_0 \) and in fact close to 1, with \( \gamma_1 \approx 1 - \lambda (1 - \gamma_0) (r - 1) / (\gamma_0 (r + 1)) \). In the next section we will see that this is true in practice.

5 Numerical simulations

We consider the ill-posed problem of estimating the second derivative function \( f(x) = g''(x) \), \( 0 \leq x \leq 1 \), given discrete noisy data \( y_i = g(x_i) + \epsilon_i \), where \( x_i = (i - 1) / n \), \( i = 1, \ldots, n \), and the errors are independent normal variates with mean 0 and variance \( \sigma^2 \). We take \( g(x) = (x^3 - x) / 6 \) so the unknown solution is simply \( f_0(x) = x \). Estimates were obtained using first order regularized of a corresponding discrete model \( y_i = (K f_c)_i + \epsilon_i \). This is the same example used in [13] and more details can be found there. Qualitatively similar results were obtained for other ill-posed integral equations and for spline smoothing. More extensive simulations of the RGCV and \( R_1 \) GCV methods will be reported elsewhere.

Our computations were carried out in MATLAB with the aid of the package Regularization Tools of Hansen [5], which is available from the Netlib library.

With \( n = 51 \) and \( \sigma = 0.001 \), the GCV method does not always give a good value of the regularization parameter [13]: for 20 replicates of the data the functions \( V(\lambda) \) are seen to have high variability for small values of \( \lambda \), and to have spurious minimizers in some cases.

For the same 20 replicates, Figure 1 shows the functions \( F_1(\lambda) \) defined in (3.5) together with \( EF_1(\lambda) \) and \( ER_1(\lambda) \) defined in (5.1). Although \( F_1(\lambda) \) does not follow \( ER_1(\lambda) \) on both sides of the minimum point of \( ER_1(\lambda) \) (marked with a + sign), it does approximate \( ER_1(\lambda) \) well in an interval to the left of the minimizer, consistent with Theorem 4.4.

Figure 1: 20 replicates of \( F_1(\lambda) \) (dotted), \( EF_1(\lambda) \) (dashed) and \( ER_1(\lambda) \) (solid)

Figure 2: 20 replicates of \( \overline{V}_1(\lambda) \) (dotted) and \( E\overline{R}_1(\lambda) + \gamma \sigma^2 \), for \( \gamma = 0.99999 \), \( \sigma = 0.001 \) (solid)

Again for the same 20 replicates, Figure 2 shows the strong robust GCV functions \( \overline{V}_1(\lambda) \) defined in (3.6) with \( \gamma = 0.99999 \), together with the shifted strong robust risk \( E\overline{R}_1(\lambda) + \gamma \sigma^2 \)
defined in (4.6). Clearly, the functions \( \nabla_1(\lambda) \) track \( ER_1(\lambda) + \gamma\sigma^2 \) in a neighbourhood of its minimizer, consistent with Theorem 4.3. Figure 2 shows that (see also Figure 3 in [13]) even though \( \gamma = 0.9999 \) is close to 1, the shape of each function \( \nabla_1(\lambda) \) is a big improvement over that of \( V(\lambda) \). For all replicates the global minimizer is close to the minimizer of \( ER_1(\lambda) \). The improvement is greater than for the function \( \nabla(\lambda) \) in the RGCV method with \( \gamma = 0.1 \) (see Figure 6 in [13]).

To assess the accuracy and reliability of the methods, GCV and \( R_1 \) GCV with \( \gamma = 0.9999 \) were used to obtain estimates \( \hat{\lambda}_V \) and \( \hat{\lambda}_{R_1} \), respectively, for each of 100 replicates of the data (again for \( n = 51 \) and \( \sigma = 0.001 \)). Then for each estimate \( \hat{\lambda} \) we computed the inefficiencies:

(a) \( I_{R_1}(\hat{\lambda}) = R(\hat{\lambda})/\min \{ R(\lambda) \} \),  
(b) \( I_{ER_1}(\hat{\lambda}) = ER(\hat{\lambda})/\min \{ ER(\lambda) \} \),
(c) \( I_{R_1}(\hat{\lambda}) = R_1(\hat{\lambda})/\min \{ R_1(\lambda) \} \),  
(d) \( I_{ER_1}(\hat{\lambda}) = ER_1(\hat{\lambda})/\min \{ ER_1(\lambda) \} \),

where \( R(\lambda) = n^{-1}\| K\hat{f}_\lambda - Kf_\epsilon \|^2 \) and

\[
R_1(\lambda) = \| K\hat{f}_\lambda - Kf_\epsilon \|^2 = \sum_{i=1}^{n} n^{-2} (K\hat{f}_\lambda - Kf_\epsilon; \hat{\phi}_i)^2 \hat{\lambda}_i^{-1}.
\]  

Of the two loss functions \( R(\lambda) \) and \( R_1(\lambda) \) above, \( R_1(\lambda) \) is the better measure of the accuracy of the regularized solution \( f_\lambda \) (see [13]).

For the GCV estimate \( \hat{\lambda} = \hat{\lambda}_V \), histograms with bin width 0.5 for the four inefficiencies above are shown in Figures 3(a), (b), (c) and (d), respectively (the same as Figure 7 in [13]). If the inefficiency is greater than or equal to 50, it is included in the bin at 50. These histograms show that for 60-80% of the replicates, the corresponding inefficiency satisfies \( 1 \leq I \leq 1.5 \), but for nearly 20% of the replicates the GCV estimate has a very large inefficiency, i.e., it produced a very poor regularized solution.

Figures 4(a), (b), (c) and (d) show the inefficiency histograms of the \( R_1 \) GCV estimate \( \hat{\lambda} = \hat{\lambda}_{R_1} \) for the same 100 replicates. Clearly, for 90-100% of the replicates, the inefficiencies in (b), (c) and (d) are less than 1.5, so the method is accurate, and there is no replicate for which any inefficiency is large, so the method is very reliable. These results are a substantial improvement on the results for GCV. They are also slightly better than those obtained by RGCV for the same 100 replicates, as shown in Figure 8 of [13].

To see the effect of the choice of \( \gamma \), Figure 5 shows the proportion of the 100 replicates for which the \( R_1 \) GCV estimate \( \hat{\lambda} \) gives inefficiency \( I_{R_1}(\hat{\lambda}) > 1.5 \) (+ symbol), \( I_{R_1}(\hat{\lambda}) > 2 \) (+ symbol) and \( I_{R_1}(\hat{\lambda}) > 4 \) (o symbol). For a better representation the proportions are plotted against \( 1 - \gamma \). Clearly all the proportions tend to decrease significantly as \( 1 - \gamma \) increases from 0 (the GCV case) up to about \( 10^{-4} \), i.e., \( \gamma = 0.9999 \), for which \( R_1 \) GCV is accurate and reliable as discussed above. As \( 1 - \gamma \) increases past \( 10^{-3} \), there is a rapid rise in the proportion of replicates for which \( I_{R_1}(\hat{\lambda}) > 1.5 \), indicating a significant loss of accuracy. These results are consistent with the remark in Section 3 that \( \gamma \) in \( R_1 \) GCV should be taken close to 1. Also the value \( 1 - \gamma = 10^{-4} \) used in \( R_1 \) GCV above is consistent with (4.13), since RGCV with \( \gamma = 0.1 \) gives a good estimate \( \lambda \in [10^{-5}, 10^{-4}] \) for this example (see [13]), and so, using \( \gamma_0 = 0.1 \) and \( r = 6 \) in (4.13), we get \( 1 - \gamma_1 \in [6 \times 10^{-5}, 6 \times 10^{-4}] \).

Acknowledgment. I thank L. Tenorio for drawing my attention to the modified GCV method and for other useful comments which significantly improved the presentation of this paper.
Figure 3: Histograms of inefficiencies (a) $I_R(\hat{\lambda})$, (b) $I_{ER}(\hat{\lambda})$, (c) $I_{Ri}(\hat{\lambda})$ and (d) $I_{ERi}(\hat{\lambda})$ for GCV estimate $\hat{\lambda} = \hat{\lambda}_V$.

Figure 4: Histograms of inefficiencies (a) $I_R(\check{\lambda})$, (b) $I_{ER}(\check{\lambda})$, (c) $I_{Ri}(\check{\lambda})$ and (d) $I_{ERi}(\check{\lambda})$ for R1GCV ($\gamma = 0.9999$) estimate $\check{\lambda} = \check{\lambda}_V$.

References

Figure 5: Proportion of R1 GCV replicates with $I_{R_1}(\hat{\lambda}) > 1.5$ (*), $I_{R_1}(\hat{\lambda}) > 2$ (+) and $I_{R_1}(\hat{\lambda}) > 4$ (○)


