Robust Generalized Cross-Validation for Choosing the Regularization Parameter

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

Abstract. Let f_λ be the regularized solution for the problem of estimating a function or vector f_0 from noisy data $y_i = L_i f_0 + \varepsilon_i$, $i = 1, \ldots, n$, where L_i are linear functionals. A prominent method for the selection of the crucial regularization parameter λ is generalized cross-validation (GCV). It is known that GCV has good asymptotic properties as $n \to \infty$ but it may not be reliable for small or medium sized n, sometimes giving an estimate that is far too small. We propose a new robust GCV method (RGCV) which chooses λ to be the minimizer of $\gamma V(\lambda) + (1-\gamma)F(\lambda)$, where $V(\lambda)$ is the GCV function, $F(\lambda)$ is an approximate average measure of the influence of each data point on f_λ, and $\gamma \in (0,1)$ is a robustness parameter. We show that for any n, RGCV is less likely than GCV to choose a very small value of λ, resulting in a more robust method. We also show that RGCV has good asymptotic properties as $n \to \infty$ for general linear operator equations with uncorrelated errors. The function $EF(\lambda)$ approximates the risk $ER(\lambda)$ for values of λ that are asymptotically a bit smaller than the minimizer of $ER(\lambda)$ (where $V(\lambda)$ may not approximate well). The "expected" RGCV estimate is asymptotically optimal as $n \to \infty$ with respect to the "robust risk" $\gamma ER(\lambda) + (1-\gamma)v(\lambda)$, where $v(\lambda)$ is the variance component of the risk, and it has the optimal decay rate with respect to $ER(\lambda)$ and stronger error criteria. The GCV and RGCV methods are compared in numerical simulations for the problem of estimating the second derivative from noisy data. The results for RGCV with $n = 51$ are consistent with the asymptotic results, and, for a large range of γ values, RGCV is more reliable and accurate than GCV.

Subject Classifications: AMS(2000) 65J20, 65J22, 45Q05, 62G08, 62J07

1 Introduction

Consider the problem of estimating a function or vector f_0 from indirect measurements

$$y_i = L_i f_0 + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where L_i are linear functionals and ε_i are random errors. An important class of examples are linear inverse problems or ill-posed operator equations $Kf(x) = g(x)$, $x \in [0,1]$, e.g. a first kind Fredholm integral equation

$$Kf(x) \;=\; \int_0^1 k(x,t) f(t)\, dt \;=\; g(x), \quad x \in [0,1], \tag{1.2}$$

for which we wish to estimate the solution f_0 from discrete noisy data $y_i = g(x_i) + \varepsilon_i$, $i = 1, \ldots, n$. Here the functionals are $L_i f = Kf(x_i)$, $i = 1, \ldots, n$. The general problem also includes a discretized operator equation or other finite dimensional linear model, in which case we have $L_i \boldsymbol{f} = K \boldsymbol{f}_i$, where $\boldsymbol{f} \in \mathbb{R}^q$, $q \le n$, and K is the $n \times q$ model or design matrix.

To obtain an approximate solution of this problem, we use the well-known method of regularization in the form (see [26])

$$\text{minimize} \quad n^{-1} \sum_{i=1}^{n} (L_i f - y_i)^2 \;+\; \lambda \|Pf\|_W^2 \tag{1.3}$$

over $f \in W$. Here $\lambda > 0$ is called the regularization parameter and $\|Pf\|_W^2$ is a roughness penalty defined by an appropriate Hilbert space W. The operator $P : W \to W$ is either the identity or an orthogonal projection with finite dimensional null space. In particular, if W is a Sobolev space of order 2 with a certain inner product and projection P, then $\|Pf\|_W^2 = \int (f''(x))^2 dx$ (see [26]).

In the special case of (1.1) where $L_i f = f(x_i)$, $x_i \in [0,1]$, we have a problem of data smoothing or fitting a curve to noisy data. In this case it is well known [26] that if $\|Pf\|_W^2 = \int (f^{(m)}(x))^2 dx$ in (1.3), then f_λ is the natural polynomial smoothing spline of degree $2m-1$.

In the case of a discrete linear model $y_i = (K\boldsymbol{f}_0)_i + \varepsilon_i$, $i = 1, \ldots, n$, where $\boldsymbol{f}_0 \in \mathbb{R}^q$, we apply regularization of the form

$$\text{minimize} \quad n^{-1} \sum_{i=1}^{n} (K\boldsymbol{f}_i - y_i)^2 \;+\; \lambda \|M\boldsymbol{f}\|^2 \tag{1.4}$$

over $\boldsymbol{f} \in \mathbb{R}^q$, with Euclidean norm $\|\cdot\|$ and suitable matrix M. Usually, either $M = I$ or $M\boldsymbol{f}$ involves first or second finite differences of $\boldsymbol{f}$. This method is also known as ridge regression (see [7]).

Under mild conditions, (1.3) and (1.4) have a unique solution f_λ, called the regularized solution (see Section 2). It is well known that to obtain a reasonable regularized solution it is important to make a good choice of λ. If λ is too small, the regularized solution is too noisy, while if λ is too large, the regularized solution is overly smooth and inaccurate.

One of the most prominent and successful methods for choosing the regularization parameter in (1.3) or (1.4) is generalized cross-validation (GCV). It is easy to use in practice and requires no knowledge of the error variance or the smoothness of the desired solution. Let $A = A(\lambda)$ be the

influence matrix defined by $A\boldsymbol{y} = \boldsymbol{L}f_\lambda$, where $\boldsymbol{L}f = (L_1 f, \ldots, L_n f)^T$. In particular, the influence matrix for the regularized solution of (1.4) is $A = K(K^TK + n\lambda M^TM)^{-1}K^T$ (see Section 2). The GCV choice of λ is the minimizer of the GCV function

$$V(\lambda) = \frac{n^{-1}\|(I-A)\boldsymbol{y}\|^2}{[n^{-1}\mathrm{tr}(I-A)]^2}, \tag{1.5}$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$. The GCV function $V(\lambda)$ can be expressed as 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 kth data point.

The GCV method was introduced by Wahba [24], and investigated further in [5], [7], [25], [15], [9], [18], [19] and [23]. See also [1] and [16] for related results on an unbiased risk criterion to choose the regularization parameter for the white noise model corresponding to the discrete data model (1.1). From these papers it is known that the GCV method has good asymptotic properties as $n \to \infty$, being asymptotically optimal in appropriate senses, and it performs well in practice for large n.

However, it is known that for small or medium sized n, the GCV method may not be reliable and can give a value of λ that is far too small (possibly even 0), corresponding to a very noisy regularized solution; see section 4.9 in [26] and [12]. In the case of spline smoothing, Wahba and Wang [27] showed that there is a nonzero probability that the GCV function has a local minimum at 0, and though this probability goes to 0 exponentially fast as $n \to \infty$, it can be significant for small n. Also, for finite n the GCV function can have several local minima, a property that was analysed by Hall and Marron [8] for kernel density estimation. Efron [6] and Kou and Efron [13] gave an intuitive geometric explanation of why GCV and the related unbiased risk criterion C_p are unstable for small n. Using this and other ideas, they developed a more stable method, called the extended exponential criterion; see also [14].

Here we propose a new general method, called robust GCV (RGCV), for choosing the regularization parameter. It will be shown that this method is more reliable than GCV for small n as well as being potentially more accurate than GCV in general.

The RGCV method is defined using the average influence $n^{-1}\sum_k \|\boldsymbol{L}f_\lambda - \boldsymbol{L}f_\lambda^{[k]}\|^2$, where $\|\boldsymbol{L}f_\lambda - \boldsymbol{L}f_\lambda^{[k]}\|^2$ is a measure of the influence of the kth data point on the regularized solution. This measure is an extension of the Cook distance [2, 3] for linear regression. Using a suitable approximation (see Section 3), the average influence simplifies to $F(\lambda) = \mu_2(\lambda)V(\lambda)$, where $\mu_2(\lambda) = n^{-1}\mathrm{tr}(A^2)$. The RGCV estimate is defined to be the minimizer of the weighted sum

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

where $\gamma \in (0, 1]$ is a robustness parameter. Clearly, when $\gamma = 1$, RGCV reduces to GCV. In the special case of data smoothing by natural polynomial smoothing splines, the same parameter choice function was derived in a different way (using a special property of the splines) by Robinson and Moyeed [22].

In Section 3 it is shown that the term $(1-\gamma)F(\lambda)$ in $\overline{V}(\lambda)$ penalizes values of λ that are close to 0. We also show that for any n, RGCV is less likely than GCV to choose a very small value of

λ (including 0), resulting in a more robust method. The smaller the value of γ the more robust is the RGCV method.

Note that if a singular value decomposition (SVD) is used to compute tr A in the GCV function $V(\lambda)$, there is little extra calculation required to compute $\mu_2(\lambda)$ in the RGCV function $\overline{V}(\lambda)$. The parameter γ enters in $\overline{V}(\lambda)$ in a very simple way and so, as described in Section 3, it is feasible to try several different values of γ to find one that gives a good choice of λ.

RGCV also has favourable properties for large n which further explain its robustness. In Section 4 we derive asymptotic results as $n \to \infty$ for the RGCV method for (1.3), with $L_i f = Kf(x_i)$ for a linear operator K. We use the same assumptions as in [18]. In particular, we assume that the errors ε_i are uncorrelated random variables, each with mean 0 and variance σ^2.

For the GCV method it is known [26, 18] that the function $EV(\lambda) - \sigma^2$ approximates the risk $ER(\lambda) = n^{-1}E\|\boldsymbol{L}f_\lambda - \boldsymbol{L}f_0\|^2$ (expected mean square prediction error) in a neighbourhood of the minimizer λ_R of $ER(\lambda)$. In Theorem 4.1 we show that $EF(\lambda)$ also approximates $ER(\lambda)$ but for values of λ that are asymptotically a bit smaller than λ_R. This is useful because by itself $V(\lambda)$ sometimes deviates significantly from $ER(\lambda) + \sigma^2$ for such values of λ.

With the assumption of uncorrelated errors, the risk can be expressed as $ER(\lambda) = b^2(\lambda) + v(\lambda)$, where $b^2(\lambda) = n^{-1}\|E\boldsymbol{L}f_\lambda - \boldsymbol{L}f_0\|^2$ is the squared bias and $v(\lambda) = n^{-1}E\|\boldsymbol{L}f_\lambda - E\boldsymbol{L}f_\lambda\|^2$ is the variance. In Theorem 4.2 we show that the minimizer $\lambda_{\overline{V}}$ of $E\overline{V}(\lambda)$ tends to minimize the "robust risk" $E\overline{R}(\lambda) = \gamma ER(\lambda) + (1-\gamma)v(\lambda)$ in that the inefficiency $E\overline{R}(\lambda_{\overline{V}})/\min E\overline{R}(\lambda) \to 1$ as $n \to \infty$. The robust risk differs from the risk $ER(\lambda)$ only in putting extra weight on the variance $v(\lambda)$. Therefore the optimal parameters with respect to the risk and robust risk have the same decay rate as $n \to \infty$, but with different coefficients depending on γ. Hence $\lambda_{\overline{V}}$ also has the optimal decay rate. As shown in Corollary 4.1, if f_0 is not too "smooth" relative to W, this decay rate is also optimal for a range of stronger error functions, including $E\|f_\lambda - f_0\|_{L^2}^2$ and $E\|f_\lambda - f_0\|_W^2$.

Section 5 describes numerical simulations with the GCV and RGCV methods for the discretized ill-posed problem of estimating the second derivative of a function $g(x)$ from noisy data $y_i = g(x_i) + \varepsilon_i$, $i = 1, \ldots, n$. To measure the accuracy of the regularized solution $\boldsymbol{f}_\lambda$ of (1.4) we use two loss functions: the mean square prediction error $R(\lambda) = n^{-1}\|K\boldsymbol{f}_\lambda - K\boldsymbol{f}_0\|^2$ and a stronger loss function $R_1(\lambda) = \|K\boldsymbol{f}_\lambda - K\boldsymbol{f}_0\|_1^2$, which behaves like a squared discrete Sobolev norm of order 1 of the error $\boldsymbol{f}_\lambda - \boldsymbol{f}_0$. The second one is the better guide to the accuracy of $\boldsymbol{f}_\lambda$. The GCV and RGCV estimates were found for 100 replicates of the data, and inefficiencies with respect to $R(\lambda)$, $ER(\lambda)$, $R_1(\lambda)$ and $ER_1(\lambda)$ were computed and displayed in histograms. Even for moderately sized n the numerical results are consistent with the results in Section 4. For $n = 51$ data points and errors ε_i with standard deviation $\sigma = 0.001$ (about 1.5% error), GCV gave a poor choice of λ in about 20% of the replicates. With $\gamma = 0.1$, the RGCV method gave a good choice for almost all the replicates. Other values of γ were also tried, and a plot of the proportion of poor choices against γ shows that for $\gamma \in [0.1, 0.5]$ the RGCV method gives less than half the number of poor choices as GCV.

In [20] we derive and investigate a whole family of robust GCV methods, ranging from the RGCV method above to a strong robust GCV method.

2 The regularized solution and spectral representation

Assume that for each $i = 1, \ldots, n$, the linear functional $L_i : W \to \mathbb{R}$ is bounded and let η_i be its representer, defined by $L_i f = (f, \eta_i)_W$ for all $f \in W$. For notational convenience, define $\boldsymbol{L} : W \to \mathbb{R}^n$ by $\boldsymbol{L}f = (L_1 f, \ldots, L_n f)^T$. It is well known [24, 26] that if $P = I$, (1.3) has the unique solution

$$f_\lambda \;=\; \boldsymbol{\eta}^T (Q + n\lambda I)^{-1} \boldsymbol{y},$$

where $\boldsymbol{\eta}^T = (\eta_1, \ldots, \eta_n)$ and Q is the $n \times n$ matrix with elements $Q_{ij} = (\eta_i, \eta_j)_W$. Then the influence matrix A, defined by $A\boldsymbol{y} = \boldsymbol{L}f_\lambda$, is $A \;=\; Q(Q + n\lambda I)^{-1}$ and the residual vector is

$$\boldsymbol{y} - A\boldsymbol{y} \;=\; n\lambda (Q + n\lambda I)^{-1} \boldsymbol{y}. \tag{2.1}$$

If $P \neq I$, let $\{\theta_j, j = 1, \ldots, m\}$, where $m << n$, be a basis for $N(P)$ and let $\xi_i = P\eta_i$, $i = 1 \ldots, n$. Define matrices $T = T_{n\times m}$ and $\Sigma = \Sigma_{n\times n}$ by $T_{ij} = L_i \theta_j$ and $\Sigma_{ij} = (\xi_i, \xi_j)_W = L_i \xi_j$. Assume that $N(\boldsymbol{L}) \cap N(P) = \{0\}$ or, equivalently, that T has full rank m. Then there exists a matrix $B = B_{(n-m)\times n}$ such that $BB^T = I_{n-m}$ and $BT = O_{(n-m)\times m}$. In this case it is well known [11, 26] that (1.3) has the unique solution

$$f_\lambda \;=\; \sum_{i=1}^{m} a_i \theta_i + \boldsymbol{\xi}^T B^T (B\Sigma B^T + n\lambda I)^{-1} B\boldsymbol{y}, \tag{2.2}$$

where $\boldsymbol{a} = (a_1, \ldots, a_m)^T$ is the unique solution of

$$T\boldsymbol{a} \;=\; \boldsymbol{y} - (\Sigma + n\lambda I) B^T (B\Sigma B^T + n\lambda I)^{-1} B\boldsymbol{y}. \tag{2.3}$$

From (2.2) and (2.3), the influence matrix A is given by

$$A\boldsymbol{y} \;=\; T\boldsymbol{a} + \Sigma B^T (B\Sigma B^T + n\lambda I)^{-1} B\boldsymbol{y} \;=\; \boldsymbol{y} - n\lambda B^T (B\Sigma B^T + n\lambda I)^{-1} B\boldsymbol{y},$$

and therefore the residual vector is

$$\boldsymbol{y} - A\boldsymbol{y} \;=\; n\lambda B^T (B\Sigma B^T + n\lambda I)^{-1} B\boldsymbol{y}. \tag{2.4}$$

Clearly, from (2.1) and (2.4), $I - A$ and A are symmetric.

The GCV function in (1.5) can be evaluated using a spectral decomposition of the influence matrix A, as shown in [25, 26]. In the case where $P = I$, from the definition of Q, clearly $n^{-1}Q$ is symmetric and positive definite, and therefore it has eigenvalues $\bar{\lambda}_i$ such that $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_n \geq 0$ (not all equal to 0) and corresponding eigenvectors $\bar{\boldsymbol{\phi}}_i$ such that $n^{-1}(\bar{\boldsymbol{\phi}}_i, \bar{\boldsymbol{\phi}}_j) = \delta_{ij}$, where $(\cdot, \cdot)$ is the Euclidean inner product on $\mathbb{R}^n$.

It is worth noting that these definitions involve the appropriate normalizations to be consistent with the limiting form of the problem. Suppose that $L_i f = Kf(x_i)$ for a bounded linear operator $K : W \to L^2(0,1)$. Assume that as $n \to \infty$, the empirical distribution function G_n of the points $x_1, \ldots, x_n$ approaches a distribution function G, and let $L^2(G)$ be $L^2(0,1)$ with inner product defined by $\int g(x)h(x)dG$. Then $\bar{\lambda}_i$ and $\bar{\boldsymbol{\phi}}_i$, $i = 1, \ldots, n$, are approximate eigenvalues and eigenfunctions (orthonormal in $L^2(G)$) of the operator KK^*, where $K^* : L^2(G) \to W$ is the adjoint of K. This is discussed further in [18, 19].

In the case where $P \neq I$, since $n^{-1}B\Sigma B^T$ is symmetric and positive definite, there exists an orthogonal matrix $U = U_{(n-m)\times(n-m)}$ such that $n^{-1}B\Sigma B^T = U\Lambda U^T$, where $\Lambda = \mathrm{diag}\{\bar{\lambda}_1, \ldots, \bar{\lambda}_{n-m}\}$ and $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_{n-m} \geq 0$ (not all equal to 0). Let $W = W_{n\times(n-m)} = B^T U$. Then $W^T W = I_{n-m}$, and from (2.4) we get $I - A = \lambda W(\Lambda + \lambda I)^{-1} W^T$. Let $\boldsymbol{w}_i$ be the ith column of W and define $\bar{\boldsymbol{\phi}}_i = \sqrt{n}\boldsymbol{w}_i$, so $n^{-1}(\bar{\boldsymbol{\phi}}_i, \bar{\boldsymbol{\phi}}_j) = \delta_{ij}$.

With these definitions, we have for both $P = I$ and $P \neq I$ that

$$A\bar{\boldsymbol{\phi}}_i \;=\; A^T\bar{\boldsymbol{\phi}}_i \;=\; [\bar{\lambda}_i/(\bar{\lambda}_i + \lambda)]\bar{\boldsymbol{\phi}}_i. \tag{2.5}$$

When $P \neq I$, the residual vector in (2.4) can be expressed as

$$\boldsymbol{y} - A\boldsymbol{y} \;=\; \lambda \sum_{i=1}^{n-m} n^{-1}(\boldsymbol{y}, \bar{\boldsymbol{\phi}}_i)\bar{\boldsymbol{\phi}}_i/(\bar{\lambda}_i + \lambda). \tag{2.6}$$

From this and the decomposition of $I - A$, we obtain

$$n^{-1}\|(I - A)\boldsymbol{y}\|^2 \;=\; \lambda^2 \sum_{i=1}^{n-m} n^{-2}(\boldsymbol{y}, \bar{\boldsymbol{\phi}}_i)^2/(\bar{\lambda}_i + \lambda)^2 \quad \text{and} \tag{2.7}$$

$$n^{-1}\mathrm{tr}(I - A) \;=\; n^{-1}\lambda \sum_{i=1}^{n-m} 1/(\bar{\lambda}_i + \lambda). \tag{2.8}$$

These expressions can be used to compute the GCV function $V(\lambda)$ in (1.5). When $P = I$, from (2.1), the same equations (2.6)-(2.8) apply but with $m = 0$.

In later sections we will use the important functions $\mu_1(\lambda) = n^{-1}\mathrm{tr}\, A$ and $\mu_2(\lambda) = n^{-1}\mathrm{tr}(A^2)$. Using the spectral decompositions above, μ_1 and μ_2 can be expressed as

$$\mu_1(\lambda) \;=\; n^{-1}\mathrm{tr}\, A \;\;=\;\; n^{-1}\left(m + \sum_{i=1}^{n-m} \bar{\lambda}_i/(\bar{\lambda}_i + \lambda)\right), \tag{2.9}$$

$$\mu_2(\lambda) \;=\; n^{-1}\mathrm{tr}(A^2) \;\;=\;\; n^{-1}\left(m + \sum_{i=1}^{n-m} [\bar{\lambda}_i/(\bar{\lambda}_i + \lambda)]^2\right) \tag{2.10}$$

if $P \neq I$, and the same form but with $m = 0$ if $P = I$. Note that $0 < \mu_1 < 1$, $0 < \mu_2 < 1$ and $\mu_2 < \mu_1$ for $\lambda > 0$. Also, by the Cauchy-Schwartz inequality, $\mu_1^2 \leq \mu_2$ for all $\lambda \geq 0$. Clearly the GCV function in (1.5) can be expressed as

$$V(\lambda) \;=\; n^{-1}\|(I - A)\boldsymbol{y}\|^2/(1 - \mu_1)^2.$$

Discrete regularization method

The fully discrete regularization problem (1.4), where K is $n \times q$, $q \leq n$, and M is $p \times q$, can be expressed as a standard least squares problem:

$$\text{minimize} \quad n^{-1}\left\| \begin{pmatrix} K \\ \sqrt{n\lambda}M \end{pmatrix} \boldsymbol{f} - \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \right\|^2$$

over $\boldsymbol{f} \in \mathbb{R}^q$. From this it is not hard to see that if $N(K) \cap N(M) = \{0\}$, there is a unique regularized solution $\boldsymbol{f}_\lambda = (K^T K + n\lambda M^T M)^{-1} K^T \boldsymbol{y}$, and the influence matrix is $A = K(K^T K + n\lambda M^T M)^{-1} K^T$. Clearly A is symmetric.

In the case where $M = I_q$, it is well known that $\boldsymbol{f}_\lambda$ and the GCV function $V(\lambda)$ can be computed using the singular value decomposition (SVD) $K = USV^T$. Here the $n \times n$ matrix U and the $q \times n$ matrix V satisfy $U^TU = V^TV = I_n$, and $S = \mathrm{diag}\{\sigma_1, \ldots, \sigma_n\}$, where the singular values σ_i are ordered such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. Define $\bar{\lambda}_i \equiv n^{-1}\sigma_i^2$, $i = 1, \ldots, n$, and $\bar{\boldsymbol{\phi}}_i \equiv \sqrt{n}\boldsymbol{u}_i$, where $\boldsymbol{u}_i$ is the ith column of U (so $n^{-1}(\bar{\boldsymbol{\phi}}_i, \bar{\boldsymbol{\phi}}_j) = \delta_{ij}$). Then it is not hard to show that (2.5) holds for all i, that the residual vector $\boldsymbol{y} - A\boldsymbol{y}$ is given by the same expression as in (2.6) but with $m = 0$, and the equations (2.7)-(2.10) apply with $m = 0$.

In the case where $M \neq I$, suppose that M is a $p \times q$ matrix with $p \leq q \leq n$. In practice, usually we also have $q - p << q$. It is known [10] that the regularized solution $\boldsymbol{f}_\lambda$ and the GCV function $V(\lambda)$ can be computed using the generalized SVD of the pair (K, M). This is defined by

$$K = U \begin{pmatrix} S & 0 \\ 0 & I_{q-p} \end{pmatrix} X^{-1}, \quad M = V(D \;\; 0)X^{-1}, \tag{2.11}$$

where both the $n \times q$ matrix U and the $p \times p$ matrix V have orthonormal columns, the $q \times q$ matrix X is nonsingular, and $S = \mathrm{diag}\{\sigma_1, \ldots, \sigma_p\}$ and $D = \mathrm{diag}\{\delta_1, \ldots, \delta_p\}$ are $p \times p$ diagonal matrices. The elements of S and D satisfy $0 \leq \sigma_1 \leq \cdots \leq \sigma_p \leq 1$, $1 \geq \delta_1 \geq \cdots \geq \delta_p > 0$ and $\sigma_i^2 + \delta_i^2 = 1$, $i = 1, \ldots, p$. The generalized singular values are the ratios σ_i/δ_i, $i = 1, \ldots, p$, and the squares $(\sigma_i/\delta_i)^2$ are the generalized eigenvalues of (K^TK, M^TM) satisfying $K^TK\boldsymbol{x}_i = (\sigma_i/\delta_i)^2 M^TM\boldsymbol{x}_i$, $i = 1, \ldots, p$, where $\boldsymbol{x}_i$ is the ith column of X. Define $\bar{\lambda}_i \equiv n^{-1}(\sigma_{p+1-i}/\delta_{p+1-i})^2$, $i = 1, \ldots, p$, (so $\bar{\lambda}_i$ are decreasing) and $\bar{\boldsymbol{\phi}}_i \equiv \sqrt{n}\boldsymbol{u}_i$, $i = 1, \ldots, q$, where $\boldsymbol{u}_i$ is the ith column of U (so $n^{-1}(\bar{\boldsymbol{\phi}}_i, \bar{\boldsymbol{\phi}}_j) = \delta_{ij}$).

Let $P_q : \mathbb{R}^n \to \mathbb{R}^n$ denote the orthogonal projection onto $\mathrm{span}\{\bar{\boldsymbol{\phi}}_1, \ldots, \bar{\boldsymbol{\phi}}_q\}$. Using the decomposition (2.11) in the expression for A, it is not hard to show that

$$A\boldsymbol{y} \;=\; \sum_{i=1}^{p} [\bar{\lambda}_i/(\bar{\lambda}_i + \lambda)] n^{-1}(\boldsymbol{y}, \bar{\boldsymbol{\phi}}_i)\bar{\boldsymbol{\phi}}_i + \sum_{i=p+1}^{q} n^{-1}(\boldsymbol{y}, \bar{\boldsymbol{\phi}}_i)\bar{\boldsymbol{\phi}}_i$$

and so (2.5) holds for $i = 1, \ldots, p$, and

$$n^{-1}\|(I - A)\boldsymbol{y}\|^2 \;=\; \lambda^2 \sum_{i=1}^{p} n^{-2}(\boldsymbol{y}, \bar{\boldsymbol{\phi}}_i)^2/(\bar{\lambda}_i + \lambda)^2 + n^{-1}\|\boldsymbol{y} - P_q\boldsymbol{y}\|^2. \tag{2.12}$$

With the same decomposition we also obtain:

$$\mu_1(\lambda) \;=\; n^{-1}\mathrm{tr}\, A \;\;=\;\; n^{-1}\left(q - p + \sum_{i=1}^{p} \bar{\lambda}_i/(\bar{\lambda}_i + \lambda)\right), \tag{2.13}$$

$$\mu_2(\lambda) \;=\; n^{-1}\mathrm{tr}(A^2) \;\;=\;\; n^{-1}\left(q - p + \sum_{i=1}^{p} [\bar{\lambda}_i/(\bar{\lambda}_i + \lambda)]^2\right). \tag{2.14}$$

3 Robust GCV method

Let $f_\lambda^{[k]}$ be the regularized solution of (1.3) or (1.4) obtained by leaving out the kth data point. We will use the "leaving-out-one" lemma (see [26]), which states that the regularized solution (of

(1.3) or (1.4)) obtained with the data $\{(x_i, y_i), i \neq k, (x_k, L_k f_\lambda^{[k]})\}$ is just $f_\lambda^{[k]}$. This implies (see [26]) that

$$y_k - L_k f_\lambda^{[k]} \;=\; (y_k - L_k f_\lambda)/(1 - a_{kk}), \tag{3.1}$$

where $a_{kk} = a_{kk}(\lambda)$ is the kth diagonal element of the influence matrix A. The formula (3.1) can be used to rewrite the ordinary cross-validation function $V_0(\lambda)$ as

$$V_0(\lambda) \;\equiv\; n^{-1} \sum_{k=1}^{n} (y_k - L_k f_\lambda^{[k]})^2 \;=\; n^{-1} \sum_{k=1}^{n} (y_k - L_k f_\lambda)^2/(1 - a_{kk})^2 \tag{3.2}$$

As described in [26], the GCV function $V(\lambda)$ is obtained from (3.2) by replacing a_{kk} by $n^{-1}\mathrm{tr}\, A$.

The "leaving-out-one" lemma also immediately gives

$$\boldsymbol{L} f_\lambda^{[k]} \;=\; A(\boldsymbol{y} - (y_k - L_k f_\lambda^{[k]})\boldsymbol{e}_k), \tag{3.3}$$

where $\boldsymbol{e}_k$ is the kth standard unit vector. Subtracting (3.3) from $\boldsymbol{L} f_\lambda = A\boldsymbol{y}$, we get

$$\boldsymbol{L} f_\lambda - \boldsymbol{L} f_\lambda^{[k]} \;=\; A(y_k - L_k f_\lambda^{[k]})\boldsymbol{e}_k,$$

so

$$L_i f_\lambda - L_i f_\lambda^{[k]} \;=\; a_{ik}(y_k - L_k f_\lambda^{[k]}) \;=\; a_{ik}(y_k - L_k f_\lambda)/(1 - a_{kk}),$$

where we have used (3.1).

Taking the square of the Euclidean norm, $\|\boldsymbol{L} f_\lambda - \boldsymbol{L} f_\lambda^{[k]}\|^2$ is a measure of the influence of the kth data point on the regularized solution. The corresponding quantity (except for a scale factor) in the case of linear regression is the Cook distance [2, 3], which is widely used for the diagnostic detection of influential observations. If the regularization parameter is too small, then, because of the resulting sensitivity, we can expect that some points will have a large influence on the regularized solution.

Define the average influence to be

$$\begin{aligned} n^{-1} \sum_{k=1}^{n} \|\boldsymbol{L} f_\lambda - \boldsymbol{L} f_\lambda^{[k]}\|^2 \;&=\; n^{-1} \sum_{k=1}^{n} \sum_{i=1}^{n} (L_i f_\lambda - L_i f_\lambda^{[k]})^2 \\ &=\; n^{-1} \sum_{k=1}^{n} \sum_{i=1}^{n} a_{ik}^2 (y_k - L_k f_\lambda)^2/(1 - a_{kk})^2. \end{aligned} \tag{3.4}$$

Now, like in defining $V(\lambda)$ from $V_0(\lambda)$, we replace a_{kk} by

$$n^{-1} \sum_{k=1}^{n} a_{kk} \;=\; n^{-1}\mathrm{tr}\, A \;\equiv\; \mu_1$$

and replace $\sum_{i=1}^{n} a_{ik}^2$ by

$$n^{-1} \sum_{k=1}^{n} \sum_{i=1}^{n} a_{ik}^2 \;=\; n^{-1}\mathrm{tr}(A^T A) \;=\; n^{-1}\mathrm{tr}(A^2) \;\equiv\; \mu_2,$$

to get the approximate average influence function

$$F(\lambda) \;\equiv\; (\mu_2/(1 - \mu_1)^2) n^{-1} \sum_{k=1}^{n} (y_k - L_k f_\lambda)^2 \;=\; (\mu_2/(1 - \mu_1)^2) n^{-1} \|(I - A)\boldsymbol{y}\|^2. \tag{3.5}$$

Note that $F(\lambda)$ is related to $V(\lambda)$ simply by $F(\lambda) = \mu_2(\lambda)V(\lambda)$.

It is reasonable to expect that for a good choice of the regularization parameter, the average influence would not be large. With this in mind, we propose the following parameter choice method (denoted RGCV): for some γ satisfying $0 < \gamma \leq 1$, choose λ to minimize

$$\overline{V}(\lambda) \;=\; \gamma V(\lambda) + (1-\gamma)F(\lambda) \;=\; \frac{\gamma + (1-\gamma)\mu_2}{(1-\mu_1)^2} n^{-1} \|(I-A)\boldsymbol{y}\|^2. \tag{3.6}$$

Note that $\overline{V}(\lambda)$ can be computed using the expressions for $n^{-1}\|(I-A)\boldsymbol{y}\|^2$, μ_1 and μ_2 in (2.7), (2.9) and (2.10), respectively, for regularization in (1.3), and the expressions in (2.12), (2.13) and (2.14), respectively, for regularization in (1.4). Clearly, there is not much more computation needed to obtain $\overline{V}(\lambda)$ than there is for $V(\lambda)$.

The RGCV method, like GCV [26], has the property of invariance under orthogonal transformations of the data. This is desirable because, if U is an orthogonal matrix and the errors are normally distributed such that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$, the problem of estimating f_0 from $\boldsymbol{y} = \boldsymbol{L}f_0 + \boldsymbol{\varepsilon}$ is equivalent to estimating f_0 from $\tilde{\boldsymbol{y}} \equiv U\boldsymbol{y} = U\boldsymbol{L}f_0 + U\boldsymbol{\varepsilon}$, since $U\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$. Applying regularization (1.3) or (1.4), since $\|U\boldsymbol{L}f - \tilde{\boldsymbol{y}}\|^2 = \|\boldsymbol{L}f - \boldsymbol{y}\|^2$, the regularized solution of the transformed problem is identical to that of the original problem. Then the influence matrix $\tilde{A}$ for the transformed problem (satisfying $\tilde{A}\tilde{\boldsymbol{y}} = U\boldsymbol{L}f_\lambda$) is just $\tilde{A} = UAU^T$. Therefore $\|(I-\tilde{A})\tilde{\boldsymbol{y}}\|^2 = \|(I-A)\boldsymbol{y}\|^2$, $\operatorname{tr}\tilde{A} = \operatorname{tr}A$ and $\operatorname{tr}(\tilde{A}^2) = \operatorname{tr}(A^2)$, so $\overline{V}(\lambda)$ is invariant under the transformation.

Clearly, when $\gamma = 1$ the RGCV method is the GCV method. The parameter γ is a robustness parameter; as γ decreases the method becomes more robust. This can be seen easily by writing the method in the equivalent form: select λ to minimize

$$(1/\gamma)\overline{V}(\lambda) \;=\; V(\lambda) + ((1-\gamma)/\gamma)F(\lambda) \;=\; [1 + ((1-\gamma)/\gamma)\mu_2(\lambda)]V(\lambda). \tag{3.7}$$

Suppose first that $P = I$ in (1.3) and $\bar{\lambda}_i > 0$ for all $i = 1, \ldots, n$. From (2.10) with $m = 0$, $\mu_2(\lambda) \to 0$ as $\lambda \to \infty$, so $(1/\gamma)\overline{V}(\lambda) \sim V(\lambda)$. Clearly $\mu_2(0) = 1$ and $\mu_2(\lambda)$ is a smooth decreasing function of λ. So as $\lambda \to 0$, we have $(1/\gamma)\overline{V}(\lambda) \sim (1/\gamma)V(\lambda) >> V(\lambda)$ for small γ. This means the term $(1-\gamma)F(\lambda)$ in $\overline{V}(\lambda)$ penalizes values of λ that are close to 0, but not large values of λ. If $P \neq I$ or $\bar{\lambda}_i > 0$ for $i = 1, \ldots, n-l$, from (2.10) the same conclusion applies so long as $m << n$ and $l << n$, which we assume is the case. For discrete regularization (1.4), we can use the same reasoning with (2.14) so long as $q - p << q$, which usually holds in practice.

Another useful perspective can be gained by taking the log in (3.7) to obtain

$$\log[(1/\gamma)\overline{V}(\lambda)] \;=\; \log V(\lambda) + \log[1 + ((1-\gamma)/\gamma)\mu_2(\lambda)]. \tag{3.8}$$

Clearly, minimizing $\log[(1/\gamma)\overline{V}(\lambda)]$ (and $\log V(\lambda)$) is equivalent to minimizing $\overline{V}(\lambda)$ (and $V(\lambda)$), and $u(\lambda) \equiv \log[1 + ((1-\gamma)/\gamma)\mu_2(\lambda)] \geq 0$ is a decreasing function of λ. Therefore in (3.8), $u(\lambda)$ is a deterministic penalty function that is summed with the random function $\log V(\lambda)$.

The following general result shows that for any n and $0 < \gamma < 1$, RGCV is less likely than GCV to choose a very small value of λ.

Theorem 3.1 *Suppose that the errors ε_i have a continuous probability distribution. Let $\overline{p} = P(\overline{V}(\lambda+h) < \overline{V}(\lambda)))$ and let $p = P(V(\lambda+h) < V(\lambda))$. For any n and $0 < \gamma < 1$, we have $\overline{p} > p$ for all $\lambda \geq 0$ and $h > 0$, and $\overline{p} - p$ increases as γ decreases.*

Proof Using the definition (1.5) of $V(\lambda)$ with (2.7) and (2.8) (or (2.12) and (2.13) for discrete regularization), it is not hard to see that for each $\lambda \geq 0$ and $h > 0$, the random variables $V(\lambda) > 0$ and $V(\lambda + h)/V(\lambda)$ have continuous distributions. Denote $z(\lambda) = \gamma + (1 - \gamma)\mu_2(\lambda)$. Then, from (3.7) and since $0 < z(\lambda + h) < z(\lambda)$, we have

$$\overline{p} \;=\; P(V(\lambda + h)/V(\lambda) < z(\lambda)/z(\lambda + h)) \;>\; P(V(\lambda + h)/V(\lambda) < 1) \;=\; p.$$

Since $\partial(z(\lambda)/z(\lambda + h))/\partial\gamma = (\mu_2(\lambda + h) - \mu_2(\lambda))/[z(\lambda + h)]^2 < 0$, clearly $\overline{p} - p$ increases as γ decreases.

□

Geometrically, it is clear that the term $(1 - \gamma)F(\lambda)$ in (3.6) has the effect of modifying the shape of $V(\lambda)$ for very small λ to better define a suitable global minimizer (see Figures 5.3 and 5.6).

We can also interpret the RGCV method intuitively as follows. If the errors ε_i are uncorrelated random variables with mean 0 and variance σ^2, it is not hard to show (see [18]) that the variance $v(\lambda) \equiv n^{-1}E\|\boldsymbol{L}f_\lambda - E\boldsymbol{L}f_\lambda\|^2$ of f_λ, as a component of the risk, is $v(\lambda) = \sigma^2\mu_2(\lambda)$. It is known [18] that under certain conditions, if $\lambda = \lambda(n) \to 0$ with the same decay rate as the minimizer of $ER(\lambda)$, then $EV(\lambda) \to \sigma^2$ as $n \to \infty$. Therefore $EF(\lambda) \sim \sigma^2\mu_2(\lambda) = v(\lambda)$, indicating that the RGCV method places extra weight on the variance of the regularized solution. This is analysed further in Section 4.2.

Clearly, the choice of the robustness parameter in the RGCV method is an important issue which needs further study. Here we just outline a simple iterative approach to select γ. For a sequence of decreasing values of γ starting with $\gamma = 1$, we compute the corresponding RGCV estimates. It would also be useful to plot $V(\lambda)$ and $\overline{V}(\lambda)$ for each value of γ to see how the graphs change. If there is a large shift in successive RGCV estimates, we can infer that $V(\lambda)$ has been modified sufficiently into $\overline{V}(\lambda)$ to produce a significantly different global minimizer. We then compute the regularized solution for the new RGCV estimate and compare it with that defined by the GCV estimate to decide which one to accept. Although it is unlikely, there may be more than one large shift if $V(\lambda)$ has several local minima. If the RGCV estimates only change continuously, we can be fairly sure that the GCV estimate, or an RGCV estimate for γ near 1, will be a good choice.

4 Asymptotic properties of the robust GCV method

4.1 Framework for asymptotic analysis

First we describe the framework for our analysis, which is the same as that in [18]. Suppose that the linear functionals $L_i : W \to \mathbb{R}$ are defined by $L_i f = Kf(x_i)$ for some bounded linear operator $K : W \to L^2(0, 1)$. Assume that for each $x \in [0, 1]$, the linear functional $W \to \mathbb{R}$, $f \to Kf(x)$ is bounded, and let η_x be its representer, so $Kf(x) = (f, \eta_x)_W$.

Assume that the empirical distribution function G_n of the points x_i, $i = 1, \ldots, n$, converges in the sup norm to a distribution function G with density bounded away from 0 and ∞. Let $L^2(G)$

denote the space $L^2(0,1)$ with inner product $(g,h)_{L^2(G)} = \int_0^1 gh\, dG$. Clearly the $L^2(G)$ norm is equivalent to the standard $L^2(0,1)$ norm. If the points x_i are equally spaced, then $G(x) = x$ and $L^2(G)$ is simply $L^2(0,1)$.

Assume that $K : W \to L^2(G)$ is $1-1$ and compact with dense range, and let $K^* : L^2(G) \to W$ be the adjoint of K. Then $K^*K : W \to W$ is compact and there is a basis $\{\psi_i\}$ for W satisfying $(\psi_i, K^*K\psi_j)_W = \delta_{ij}$, and eigenvalues τ_i satisfying $P\psi_i = \tau_i K^*K\psi_i$, with $0 \le \tau_1 \le \tau_2 \le \cdots$ and $\tau_i \to \infty$. If K is an integral operator with kernel $k(x,t)$ as in (1.2) and W is a reproducing kernel Hilbert space with kernel $R(x,t)$, then under general conditions (see [18]) the eigenvalues $\omega_1 \ge \omega_2 \ge \cdots \ge 0$ of the integral operator with kernel $kRk^*(x,t)$ satisfy $c_1\omega_i \le \tau_i^{-1} \le c_2\omega_i$ for all i. This is useful in determining the growth rate of τ_i.

To describe the "smoothness" class of f_0, we use the family of Hilbert spaces W_β as in [18] with inner product

$$(f,v)_\beta \;=\; \sum_{i=1}^{\infty}(1+\tau_i)^\beta (f, K^*K\psi_i)_W (v, K^*K\psi_i)_W.$$

It is shown in [18] that $W_1 = W$ with equivalent norms. If W is L^2 or a Sobolev space and K is a convolution integral operator with eigenvalues decaying according to a power law, then the spaces W_β can be identified as fractional Sobolev spaces, including L^2; see [18].

We now state the main assumptions in this section. These are the same assumptions used in [18] for the asymptotic analysis of GCV. For convenience we will write $a_n \approx b_n$ if there exist positive constants c_1 and c_2 such that $c_1 b_n \le a_n \le c_2 b_n$. We will also write $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$, and $a_n \lesssim b_n$ if there exists a positive constant c such that $a_n \le c b_n$.

Assumption 4.1 The errors ε_i are uncorrelated random variables with mean $E\varepsilon_i = 0$ and variance $E\varepsilon_i^2 = \sigma^2$.

Assumption 4.2

(a) The operator $K : W \to L^2$ is 1–1, bounded and compact, and $K(W)$ is dense in L^2.

(b) $P : W \to W$ is an orthogonal projection with $\dim N(P) < \infty$.

(c) There exists $r > 1$ such that $\tau_i \approx i^r$ for $i > m$.

Assumption 4.3

(a) For each $x \in [0,1]$ the functional $W \to \mathbb{R}$, $f \to Kf(x)$ is bounded.

(b) For all n sufficiently large, $N(\boldsymbol{L}) \cap N(P) = \{0\}$.

Assumption 4.4 For the kernel $q(x,t) = (\eta_x, \eta_t)_W$, there exists $\bar{q}$ such that $q(x,x) \le \bar{q}$ for all $x \in [0,1]$.

Assumption 4.5 There exists $s \in (0, 1-1/r)$, $\{\rho_1, \ldots, \rho_J\} \subseteq [0,s]$ and a sequence $d_n \to 0$ such that for all $f, v \in W$

$$|(Kf, Kv)_{L^2(G)} - n^{-1}\sum_{i=1}^{n} Kf(x_i)Kv(x_i)| \;\le\; d_n \sum_{j=1}^{J} \|f\|_{\rho_j}\|v\|_{s-\rho_j}.$$

This assumption defines the order of approximation of the integral $(Kf, Kv)_{L^2(G)}$ by the quadrature formula $n^{-1}\sum_{i=1}^n Kf(x_i)Kv(x_i)$, assuming that f and g have a certain "smoothness" de-

termined by s. If the x_i are equally spaced, then under suitable conditions, we have $d_n = O(n^{-1})$ (see [21]).

The asymptotic analysis of the RGCV method depends crucially on the asymptotic behaviour of the functions $\mu_1(\lambda)$ and $\mu_2(\lambda)$ defined in (2.9) and (2.10). The following estimates of $\mu_1(\lambda)$ and $\mu_2(\lambda)$ were derived in Theorems 4.1 and 4.3 of [18], respectively. If Assumptions 4.2–4.5 hold and $\alpha_n \to 0$ as $n \to \infty$ such that $d_n^2 \alpha_n^{-(s+1)} \to 0$, then

$$\mu_1(\lambda) \;\approx\; n^{-1} D(\lambda; -1/r, -1) \tag{4.1}$$

$$\mu_2(\lambda) \;\approx\; n^{-1} D(\lambda; -1/r, -2), \tag{4.2}$$

uniformly in $\lambda \in [\alpha_n, \infty)$, where $D(\lambda; a, b) \equiv \lambda^a$, if $\lambda \le 1$, and $D(\lambda; a, b) \equiv \lambda^b$, if $\lambda > 1$.

Under the above assumptions, the risk $ER(\lambda) = n^{-1} E\|\boldsymbol{L} f_\lambda - \boldsymbol{L} f_0\|^2$ and $E\|f_\lambda - f_0\|_W^2$ have a known asymptotic behaviour as $n \to \infty$. For $E\|f_\lambda - f_0\|_W^2$ it was derived in [4, 17, 21] and for $ER(\lambda)$ in [18]. To obtain an estimate of the risk, it is decomposed using Assumption 4.1 into a squared bias and variance term as

$$ER(\lambda) \;=\; n^{-1}\|E\boldsymbol{L} f_\lambda - \boldsymbol{L} f_0\|^2 + n^{-1} E\|\boldsymbol{L} f_\lambda - E\boldsymbol{L} f_\lambda\|^2 \;=\; b^2(\lambda) + v(\lambda), \tag{4.3}$$

where $b^2(\lambda) = n^{-1}\|(I - A)\boldsymbol{L} f_0\|^2$ is the squared bias and $v(\lambda) = \sigma^2 \mu_2(\lambda)$ is the variance. From Theorem 4.5 in [18], under Assumptions 4.1–4.5, if $f_0 \in W_\beta$, $\beta \ge s$, and $\alpha_n \to 0$ as $n \to \infty$ in such a way that

$$d_n^2 \alpha_n^{-(s+1/r)} \to 0, \quad \text{if } \; s \le \beta \le 2, \quad \text{and} \quad d_n^2 \alpha_n^{-(s+1/r+\beta-2)} \to 0, \quad \text{if } \; \beta > 2,$$

then

$$\min\{1, \lambda^2\}\|\bar{P} g\|_{L^2(G)}^2 \;\lesssim\; b^2(\lambda) \;\lesssim\; \begin{cases} \min\{1, \lambda^\beta\}\|f_0\|_\beta^2, & \text{if } \beta < 2, \\ \min\{1, \lambda^2\}\|f_0\|_2^2, & \text{if } \beta \ge 2, \end{cases} \tag{4.4}$$

uniformly in $\lambda \in [\alpha_n, \infty)$. Here $g = K f_0$ and $\overline{P} : L^2(G) \to L^2(G)$ is the orthogonal projection onto $K(N(P))^\perp$.

Using (4.2) and (4.4) in (4.3) gives an estimate of $ER(\lambda)$, and minimizing the upper bound yields Corollary 4.1 in [18] which we restate below.

Proposition 4.1 *Suppose that Assumptions 4.1–4.5 hold, $f_0 \in W_\beta$, $\beta \ge s$, and $\alpha_n \to 0$ as $n \to \infty$ in such a way that*

$$d_n^2 \alpha_n^{-(s+1)} \to 0, \quad \textit{if } \; s \le \beta \le 3 - 1/r, \quad \textit{and} \quad d_n^2 \alpha_n^{-(s+1/r+\beta-2)} \to 0, \quad \textit{if } \; \beta > 3 - 1/r.$$

Define

$$\lambda^* = \begin{cases} (\sigma^2 n^{-1})^{r/(\beta r+1)}, & s \le \beta < 2, \\ (\sigma^2 n^{-1})^{r/(2r+1)}, & \beta \ge 2, \end{cases} \tag{4.5}$$

and assume that $\lambda^* \ge \alpha_n$. Then the minimum over $\lambda \ge \alpha_n$ of the upper bound on $ER(\lambda)$ occurs at $\lambda \approx \lambda^*$ and

$$\min_{[\alpha_n, \infty)} ER(\lambda) \;\le\; ER(\lambda^*) \;\approx\; \begin{cases} (\sigma^2 n^{-1})^{\beta r/(\beta r+1)}, & s \le \beta < 2, \\ (\sigma^2 n^{-1})^{2r/(2r+1)}, & \beta \ge 2. \end{cases}$$

In addition, let $\lambda_R = \lambda_R(n)$ minimize $ER(\lambda)$ over $\lambda \geq \alpha_n$. If $f_0 \in W_\beta$, $\beta \geq 2$, then $\lambda_R \approx \lambda^*$ and $ER(\lambda_R) \approx ER(\lambda^*)$.

4.2 Asymptotic behaviour of RGCV as $n \to \infty$

Before examining the RGCV method, first we review a result about the GCV function $V(\lambda)$. It is known that, under certain assumptions, the function $EV(\lambda) - \sigma^2$ tracks the function $ER(\lambda)$ in a neighbourhood of the minimizer λ_R of $ER(\lambda)$ (see [24, 26, 18]). In fact we have

$$\frac{|EV(\lambda) - \sigma^2 - ER(\lambda)|}{ER(\lambda)} \;\leq\; h_V(\lambda) \;\equiv\; \frac{2\mu_1 + \mu_1^2/\mu_2}{(1-\mu_1)^2},$$

and, under Assumptions 4.2–4.5, if $\lambda \to 0$ as $n \to \infty$ such that $d_n^2 \lambda^{-(s+1)} \to 0$ and $n^{-1}\lambda^{-1/r} \to 0$, then $h_V(\lambda) \approx n^{-1}\lambda^{-1/r} \to 0$ as $n \to \infty$. From Theorem 5.1 in [18], $\lambda = \lambda_R$ satisfies the condition $n^{-1}\lambda^{-1/r} \to 0$.

We now show that the minimizer of $E\overline{V}(\lambda)$ will not be far away from the minimizer of the risk $ER(\lambda)$. Since $\overline{V}(\lambda) = \gamma V(\lambda) + (1-\gamma)F(\lambda)$, the minimizer of $\overline{V}(\lambda)$ will be shifted away from the minimizer of $V(\lambda)$ through the effect of $F(\lambda)$. The next result shows that $EF(\lambda)$ also tracks $ER(\lambda)$ but for values of λ that are asymptotically a bit smaller than λ^* in (4.5). This is important because $V(\lambda)$ sometimes deviates significantly from $ER(\lambda) + \sigma^2$ for such values of λ.

Theorem 4.1 *Suppose that Assumptions 4.1–4.5 hold, $f_0 \in W_\beta$, $\beta \geq s$, and $\alpha_n \to 0$ as $n \to \infty$ in the same way as in Proposition 4.1. If $\lambda = \lambda(n) \to 0$ as $n \to \infty$ such that $\lambda \geq \alpha_n$, $n\lambda^{1/r} \to \infty$ and either $\lambda^\beta(n\lambda^{1/r}) \to 0$ if $s \leq \beta < 2$ or $\lambda^2(n\lambda^{1/r}) \to 0$ if $\beta \geq 2$, then*

$$EF(\lambda) \;=\; ER(\lambda)(1 + o(1)).$$

Note that $\lambda/\lambda^* \to 0$ as $n \to \infty$ since $\lambda^{*\beta}(n\lambda^{*1/r}) \approx 1$ if $s \leq \beta < 2$, and $\lambda^{*2}(n\lambda^{*1/r}) \approx 1$ if $\beta \geq 2$.

Proof Using Assumption 4.1 and the spectral decomposition of A, we have

$$EF(\lambda) \;=\; (\mu_2/(1-\mu_1)^2) E n^{-1}\|(I-A)\boldsymbol{y}\|^2 \;=\; (\mu_2/(1-\mu_1)^2)(b^2 + \sigma^2(1 - 2\mu_1 + \mu_2)),$$

where $b^2 = b^2(\lambda) = n^{-1}\|(I-A)\boldsymbol{L}f_0\|^2$ is the squared bias. Then, from (4.3) and since $\mu_1 \to 0$, we obtain

$$\begin{aligned} |EF(\lambda) - ER(\lambda)|/ER(\lambda) \;&\leq\; |EF(\lambda) - ER(\lambda)|/(\sigma^2\mu_2) \\ &=\; b^2|1 - \mu_2/(1-\mu_1)^2|/(\sigma^2\mu_2) \;+\; \mu_2(1 - \mu_1^2/\mu_2)/(1-\mu_1)^2. \end{aligned}$$

Now, using (4.1), (4.2), and (4.4), we estimate the first term as

$$b^2|1 - \mu_2/(1-\mu_1)^2|/(\sigma^2\mu_2) \;\approx\; b^2(\lambda)(1 - O(n^{-1}\lambda^{-1/r}))/(\sigma^2 n^{-1}\lambda^{-1/r}) \;\to\; 0.$$

Using (4.1) and (4.2), the second term is estimated as

$$\mu_2(1 - \mu_1^2/\mu_2)/(1-\mu_1)^2 \;\approx\; n^{-1}\lambda^{-1/r}(1 - O(n^{-1}\lambda^{-1/r})) \;\to 0.$$

The result follows.

□

An important result about the GCV method is that the "expected" estimate is asymptotically optimal with respect to the risk, i.e. under certain conditions, there is a sequence of minimizers $\lambda_V = \lambda_V(n)$ of $EV(\lambda)$ such that the inefficiency

$$ER(\lambda_V)/\min ER(\lambda) \;\to\; 1$$

as $n \to \infty$ (see [18]).

We now derive a corresponding result for the RGCV method. Define the "robust risk" $E\overline{R}(\lambda)$ to be

$$E\overline{R}(\lambda) \;=\; \gamma ER(\lambda) + (1-\gamma)v(\lambda) \;=\; \gamma b^2(\lambda) + v(\lambda), \tag{4.6}$$

where $v(\lambda) = \sigma^2\mu_2(\lambda)$ is the variance. Clearly, when $\gamma = 1$, $E\overline{R}(\lambda) = ER(\lambda)$, the ordinary risk. But with $0 < \gamma < 1$, the robust risk $E\overline{R}(\lambda)$ is a risk function that, compared to $ER(\lambda)$, places more weight on controlling the variance of the regularized solution relative to the squared bias $b^2(\lambda)$.

From (3.6), using Assumption 4.1 and the spectral decomposition of A, we have

$$E\overline{V}(\lambda) \;=\; (\gamma + (1-\gamma)\mu_2)(b^2 + \sigma^2(1 - 2\mu_1 + \mu_2))/(1-\mu_1)^2. \tag{4.7}$$

Using the expressions in (4.6) and (4.7), and rearranging, we obtain

$$\frac{E\overline{R}(\lambda) + \gamma\sigma^2 - E\overline{V}(\lambda)}{E\overline{R}(\lambda)} \;=\; -\frac{\mu_1(2-\mu_1) + ((1-\gamma)/\gamma)\mu_2}{(1-\mu_1)^2} + \frac{\gamma\sigma^2(\mu_1 + ((1-\gamma)/\gamma)\mu_2)^2}{(\gamma b^2 + \sigma^2\mu_2)(1-\mu_1)^2}.$$

Since $\mu_1 < 1$, this gives the bound

$$\frac{|E\overline{R}(\lambda) + \gamma\sigma^2 - E\overline{V}(\lambda)|}{E\overline{R}(\lambda)} \;\le\; h(\lambda) \;\equiv\; \frac{2\mu_1 + ((1-\gamma)/\gamma)\mu_2 + \gamma(\mu_1 + ((1-\gamma)/\gamma)\mu_2)^2/\mu_2}{(1-\mu_1)^2}. \tag{4.8}$$

From the estimates in (4.1) and (4.2), if $\lambda \to 0$ as $n \to \infty$ such that $d_n^2\lambda^{-(s+1)} \to 0$ and $n^{-1}\lambda^{-1/r} \to 0$, then

$$h(\lambda) \;\approx\; (2 + (1-\gamma)/\gamma + \gamma(1 + (1-\gamma)/\gamma)^2)n^{-1}\lambda^{-1/r}/(1 - n^{-1}\lambda^{-1/r})^2 \;\to\; 0.$$

Geometrically, this means that, for a certain range of λ, the graph of $E\overline{V}(\lambda) - \gamma\sigma^2$ tracks the graph of $E\overline{R}(\lambda)$, in the same way that $EV(\lambda) - \sigma^2$ tracks $ER(\lambda)$ for the GCV method.

Using (4.8), we obtain the following result in the same way as for Lemma 5.1 in [18] and Theorem 4.2 in [5].

Lemma 4.1 *Let $a \ge 0$ and let $\lambda_{\overline{R}}$ minimize $E\overline{R}(\lambda)$ over $\lambda \ge a$. There exists a minimizer $\lambda_{\overline{V}}$ of $E\overline{V}(\lambda)$ over $\lambda \ge a$ such that if $h(\lambda_{\overline{V}}) < 1$, then*

$$\frac{E\overline{R}(\lambda_{\overline{V}})}{E\overline{R}(\lambda_{\overline{R}})} \;\le\; \frac{1 + h(\lambda_{\overline{R}})}{1 - h(\lambda_{\overline{V}})} \;=\; 1 + \frac{h(\lambda_{\overline{V}}) + h(\lambda_{\overline{R}})}{1 - h(\lambda_{\overline{V}})}.$$

The next result shows that the "expected" RGCV estimate is asymptotically optimal with respect to the robust risk $E\overline{R}(\lambda)$.

Theorem 4.2 *Suppose that Assumptions 4.1–4.5 hold, $f_0 \in W_\beta$, $\beta \geq s$, and $\alpha_n \to 0$ as $n \to \infty$ in the same way as in Proposition 4.1, and also $\alpha_n \leq \lambda^*$. Let $\lambda_{\overline{R}} = \lambda_{\overline{R}}(n)$ minimize $E\overline{R}(\lambda)$ over $\lambda \geq \alpha_n$. Then there exists a sequence $\lambda_{\overline{V}} = \lambda_{\overline{V}}(n)$ of minimizers of $E\overline{V}(\lambda)$ such that as $n \to \infty$*

$$0 \;\leq\; \frac{E\overline{R}(\lambda_{\overline{V}})}{E\overline{R}(\lambda_{\overline{R}})} - 1 \;\to\; 0.$$

Proof The proof is similar to the proof of Theorem 5.1 in [18]. First consider the case where $f_0 \in N(P)$. Then, from (2.4), $\boldsymbol{L}f_0 - A\boldsymbol{L}f_0 = 0$ and so $b^2(\lambda) = 0$. Hence, from (4.6) and (4.7), we have $E\overline{R}(\lambda) = \sigma^2\mu_2$ and

$$E\overline{V}(\lambda) = \sigma^2(\gamma + (1-\gamma)\mu_2)(1 - 2\mu_1 + \mu_2)/(1-\mu_1)^2.$$

Clearly $E\overline{R}(\lambda) = \sigma^2\mu_2$ is minimized at $\lambda = \infty$. From the argument in [5], p. 389, $(1 - 2\mu_1 + \mu_2)/(1-\mu_1)^2$ is minimized at $\lambda = \infty$, and, because $\mu_2(\lambda)$ is a decreasing function of λ, $E\overline{V}(\lambda)$ is also minimized at $\lambda = \infty$. Thus the result holds.

In the general case where $f_0 \notin N(P)$, we use Lemma 4.1 with $a = \alpha_n$ and show that $h(\lambda_{\overline{R}}) \to 0$ and $h(\lambda_{\overline{V}}) \to 0$. It is easy to see that the estimates for $ER(\lambda)$ in Proposition 4.1 also apply to $E\overline{R}(\lambda)$. Hence, from (4.6) we have, as $n \to \infty$,

$$\gamma b^2(\lambda_{\overline{R}}) \;\leq\; E\overline{R}(\lambda_{\overline{R}}) \;\leq\; E\overline{R}(\lambda^*) \;\to\; 0.$$

Therefore the lower bound in (4.4) implies that $\lambda_{\overline{R}} \to 0$. Similarly, from (4.2) and (4.6),

$$\sigma^2 n^{-1}\lambda_{\overline{R}}^{-1/r} \;\approx\; v(\lambda_{\overline{R}}) \;\leq\; E\overline{R}(\lambda_{\overline{R}}) \;\to\; 0$$

and so $n^{-1}\lambda_{\overline{R}}^{-1/r} \to 0$. Hence, using (4.8), (4.1) and (4.2), we get $h(\lambda_{\overline{R}}) \approx n^{-1}\lambda_{\overline{R}}^{-1/r} \to 0$. From (4.7) and since $\mu_1^2 \leq \mu_2$, it follows that $0 \leq E\overline{V}(\lambda) - \gamma\sigma^2$ for all $\lambda \geq 0$. Now for a minimizer $\lambda_{\overline{V}}$ of $E\overline{V}(\lambda)$, because from (4.8)

$$0 \;\leq\; E\overline{V}(\lambda_{\overline{V}}) - \gamma\sigma^2 \;\leq\; E\overline{V}(\lambda_{\overline{R}}) - \gamma\sigma^2 \;\leq\; E\overline{R}(\lambda_{\overline{R}})(1 + h(\lambda_{\overline{R}})),$$

we have $E\overline{V}(\lambda_{\overline{V}}) - \gamma\sigma^2 \to 0$. This implies, using (4.7) and $\mu_1^2 \leq \mu_2$, that

$$\gamma b^2(\lambda_{\overline{V}}) \;\leq\; \gamma b^2(\lambda_{\overline{V}})/(1-\mu_1)^2 \;\leq E\overline{V}(\lambda_{\overline{V}}) - \gamma\sigma^2 \;\to 0$$

so $\lambda_{\overline{V}} \to 0$. Also, using (4.2), (4.1) and (4.7), we have

$$\gamma\sigma^2 n^{-1}\lambda_{\overline{V}}^{-1/r} \;\approx\; \gamma\sigma^2\mu_2(\lambda_{\overline{V}}) \;\approx\; \gamma\sigma^2\mu_2(1 - \mu_1^2/\mu_2)/(1-\mu_1)^2 \;\leq\; E\overline{V}(\lambda_{\overline{V}}) - \gamma\sigma^2 \;\to\; 0$$

and so $n^{-1}\lambda_{\overline{V}}^{-1/r} \to 0$. Hence, using (4.1), (4.2) and (4.8), we get $h(\lambda_{\overline{V}}) \approx n^{-1}\lambda_{\overline{V}}^{-1/r} \to 0$. The result follows from Lemma 4.1.

□

From (4.6) it is clear that the minimizers $\lambda_{\overline{R}}$ and λ_R of $E\overline{R}(\lambda)$ and $ER(\lambda)$, respectively, have the same decay rate as $n \to \infty$, but different constant coefficients depending on γ. It is shown in Corollary 5.1 in [18] that if $f_0 \in W_\beta$, $\beta \geq 2$, then there are minimizers λ_V of $EV(\lambda)$, λ_R of $ER(\lambda)$ and λ_L of $E\|Kf_\lambda - Kf_0\|^2_{L^2(G)}$ such that $\lambda_V \approx \lambda_R \approx \lambda_L \approx \lambda^*$, where λ^* is given in (4.5).

From Theorem 4.2 we can conclude that the RGCV estimate $\lambda_{\overline{V}}$ has the same optimal rate, i.e. $\lambda_{\overline{V}} \approx \lambda_{\overline{R}} \approx \lambda^*$.

In Theorem 5.3 in [18] it is shown that if $f_0 \in W_2$ but $f_0 \notin W_{2+\delta}$, $\delta > 0$, then the GCV estimate λ_V also has the optimal decay rate for the stronger error functions $E\|f_\lambda - f_0\|_\rho^2$, $0 \le \rho \le 2$. Note this includes the error function $E\|f_\lambda - f_0\|_1^2 \approx E\|f_\lambda - f_0\|_W^2$, which is stronger than the L^2 error function $E\|f_\lambda - f_0\|_{L^2}^2$ if W is a Sobolev space. On the other hand (also from Theorem 5.3 in [18]), if $f_0 \in W_\beta$, $\beta > 2$, then for any $\rho > 0$, λ_V does not have the optimal decay rate for $E\|f_\lambda - f_0\|_\rho^2$; it decays to 0 too quickly. Hence for GCV to achieve the optimal rate for a fixed error function, say $E\|f_\lambda - f_0\|_{L^2}^2$, one should try to choose W in (1.3) such that the smoothness of f_0 is not too large relative to W. Since $\lambda_{\overline{V}} \approx \lambda_V$, the same conclusions apply to RGCV.

We summarize this in the following corollary.

Corollary 4.1 *Assume the conditions of Theorem 4.2 are satisfied. Let λ_ρ minimize $E\|f_\lambda - f_0\|_\rho^2$ over $\lambda \ge \alpha_n$ and let $\lambda_{\overline{V}}$ be the sequence of minimizers of $E\overline{V}(\lambda)$ defined in Theorem 4.2. If $f_0 \in W_2$, then for any $0 \le \rho \le 2$, we have $\lambda_{\overline{V}} \approx \lambda_\rho$ as $n \to \infty$ and*

$$E\|f_{\lambda_{\overline{V}}} - f_0\|_\rho^2 \,/\, E\|f_{\lambda_\rho} - f_0\|_\rho^2 \;=\; O(1).$$

If $f_0 \in W_\beta$, $\beta > 2$, then for any $\rho > 0$, we have $\lambda_{\overline{V}}/\lambda_\rho \to 0$ as $n \to \infty$ and

$$E\|f_{\lambda_{\overline{V}}} - f_0\|_\rho^2 \,/\, E\|f_{\lambda_\rho} - f_0\|_\rho^2 \;\to\; \infty.$$

5 Numerical simulations

To illustrate the methods and results of the previous sections, we consider the ill-posed problem of estimating the second derivative function $f_0(x) = g''(x)$, $0 \le x \le 1$, from discrete noisy data $y_i = g(x_i) + \varepsilon_i$, $i = 1, \ldots, n$, with $g(0) = g(1) = 0$. It is not hard to show that this problem is equivalent to solving the first kind Fredholm integral eqution $\int_0^1 k(x,t) f(t)\, dt \;=\; g(x)$, where

$$k(x,t) \;=\; \begin{cases} x(t-1), & x < t, \\ t(x-1), & x \ge t. \end{cases}$$

To solve the problem, we discretize the integral equation and apply the regularization method (1.4) for suitable matrix M. Using the trapezoidal rule with nodes equal to the abscissa values x_i, $i = 1, \ldots, n$, we have

$$\int_0^1 k(x_i,t) f(t)\, dt \;\approx\; \sum_{j=1}^n w_j k(x_i, x_j) f(x_j),$$

where w_j are the appropriate weights, and define the $n \times n$ matrix $K = [K_{ij}] = [w_j k(x_i, x_j)]$. For simplicity we take uniform points $x_i = (i-1)/(n-1)$, $i = 1, \ldots, n$, and let $g(x) = (x^3 - x)/6$ so $f_0(x) = x$. The data were generated using $y_i = (K\boldsymbol{f}_0)_i + \varepsilon_i$, where $\boldsymbol{f}_0 = [f_0(x_1), \ldots, f_0(x_n)]^T$ and ε_i is a pseudo-random normal variate with mean 0 and standard deviation σ. The $n \times n$ matrix M was defined by $M\boldsymbol{f}_1 = f_1$ and $M\boldsymbol{f}_i = f_i - f_{i-1}$, $i = 2, \ldots, n$, which is a (scaled) discrete approximation of the first derivative operator $f \to f'$ if $f(0) = 0$.

Our computations were carried out in MATLAB with the aid of the package Regularization Tools of Hansen [10], which is available from the Netlib library. The regularized solution $\boldsymbol{f}_\lambda = (K^T K + \lambda M^T M)^{-1} K^T \boldsymbol{y}$ was computed in this package using the generalized SVD of the pair (K, M) in (2.11). Note that since M is invertible, the generalized eigenvalues $\bar{\lambda}_i$, which satisfy $n^{-1} K^T K \bar{\boldsymbol{\phi}}_i = \bar{\lambda}_i M^T M \bar{\boldsymbol{\phi}}_i$, $i = 1, \ldots, n$, are the eigenvalues of $n^{-1}(KM^{-1})^T KM^{-1} = n^{-1}(M^T)^{-1} K^T K M^{-1}$, and also the eigenvalues of $n^{-1} KM^{-1}(KM^{-1})^T = n^{-1} K(M^T M)^{-1} K^T$.

The matrix $n^{-1} K(M^T M)^{-1} K^T$ is a scaled discrete representation of the integral operator with kernel $kGG^*k^*(x,t)$, where G is the Green's function of the derivative operator $f \to f'$ with $f(0) = 0$. Since this kernel is a Green's function for the sixth derivative operator, its eigenvalues decay like i^{-6}, corresponding to the value $r = 6$ in Assumption 4.2(c). A logarithmic plot of the computed eigenvalues $\bar{\lambda}_i$ shows the same decay rate of i^{-6} for $i = 1, \ldots, n-2$. Since the first and last columns of K are zero vectors, then $\operatorname{rank} K = n - 2$ and so $\bar{\lambda}_{n-1} = \bar{\lambda}_n = 0$.

The asymptotic results in Section 4 do not apply directly to this discretized regularization problem. However, since $K\boldsymbol{f}_i$ is a consistent approximation of $\int_0^1 k(x_i, t) f(t)\, dt$ and $n^{-1}\|(n-1)^{-1} M \boldsymbol{f}\|^2$ is a consistent approximation of $\|Pf\|_W^2 \equiv \int_0^1 (f'(t))^2\, dt$, with $f(0) = 0$, it is reasonable to expect that for sufficiently large n, the estimates in Theorems 4.1 and 4.2 will hold approximately for the discrete regularized solution $\boldsymbol{f}_\lambda$.

First we consider the GCV method. For a particular data set of size $n = 51$ with $\sigma = 0.001$, shown in Figure 5.1, the GCV method yields a good value of the regularization parameter ($\lambda = 1.076 \times 10^{-5}$) leading to the good regularized solution shown in Figure 5.2.

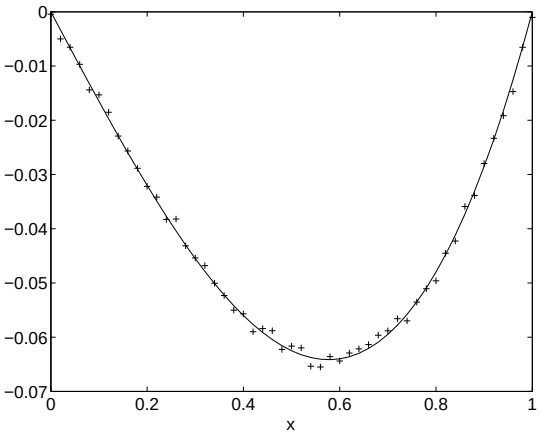

Figure 5.1: $g(x) = (x^3 - x)/6$ and data (x_i, y_i), $i = 1, \ldots, 51$

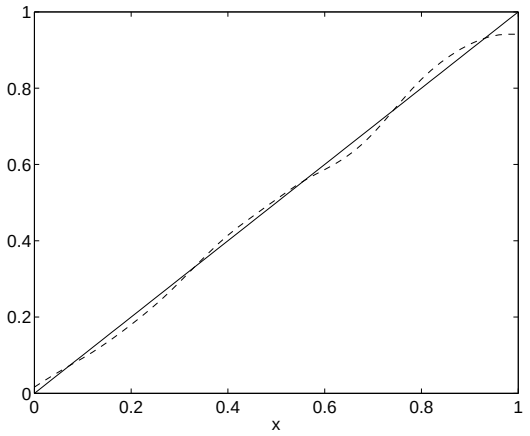

Figure 5.2: $f(x) = x$ and $f_\lambda(x)$ for good GCV estimate λ

However, GCV does not always give a good estimate. To illustrate the lack of reliability of GCV, we use 20 replicates of the data (each with different error vector but the same $\sigma = 0.001$) and plot $V(\lambda)$ together with $ER(\lambda) + \sigma^2$ for $n = 51$ in Figure 5.3 and for $n = 101$ in Figure 5.4. Here $ER(\lambda) = n^{-1} E\|K\boldsymbol{f}_\lambda - K\boldsymbol{f}_0\|^2$. As expected from the asymptotic results for GCV discussed in Section 4, in both figures the GCV functions $V(\lambda)$ track $ER(\lambda) + \sigma^2$ in a neighbourhood of the minimum of $ER(\lambda)$ and near this minimum their variability is low, as shown asymptotically

in [19]. But, clearly, for $n = 51$ in Figure 5.3, there is high variability in the GCV functions for smaller values of λ and, for several replicates, there are spurious minimizers which produce very inaccurate noisy regularized solutions. This behaviour is not nearly as pronounced for $n = 101$ in Figure 5.4. Note also that for both $n = 51$ and $n = 101$ some of the functions $V(\lambda)$ have an oscillating nature with several local minima, which is consistent with the results in [8].

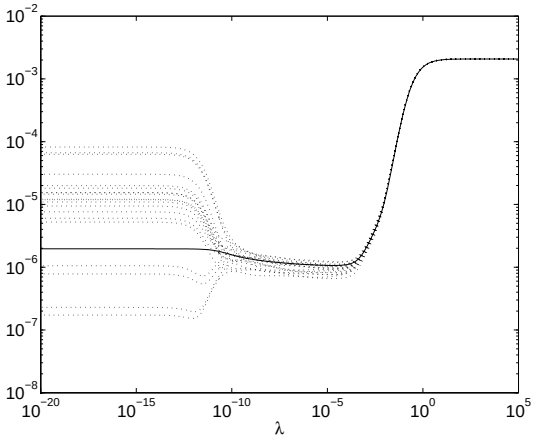

Figure 5.3: 20 replicates of $V(\lambda)$ (dotted) for $n = 51$ and $ER(\lambda) + \sigma^2$, $\sigma = 0.001$ (solid)

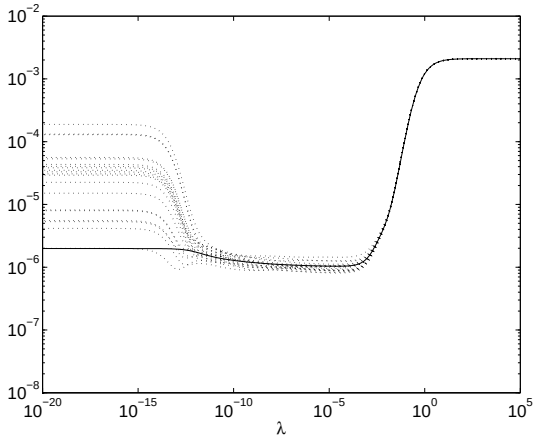

Figure 5.4: 20 replicates of $V(\lambda)$ (dotted) for $n = 101$ and $ER(\lambda) + \sigma^2$, $\sigma = 0.001$ (solid)

For $n = 51$ and the same 20 replicates of the data, Figure 5.5 shows the functions $F(\lambda)$ defined in (3.5), together with $EF(\lambda)$ and $ER(\lambda)$. Note that, although $EF(\lambda)$ does not follow $ER(\lambda)$ closely on both sides of the minimum point of $ER(\lambda)$ (marked with a + symbol), it does approximate $ER(\lambda)$ well in an interval to the left of the minimizer, consistent with Theorem 4.1.

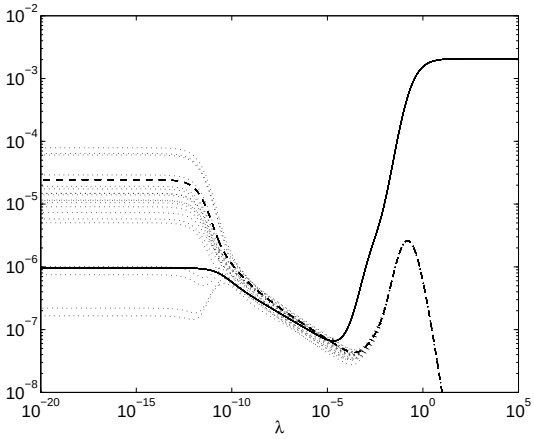

Figure 5.5: 20 replicates of $F(\lambda)$ (dotted), $EF(\lambda)$ (dashed) and $ER(\lambda)$ (solid)

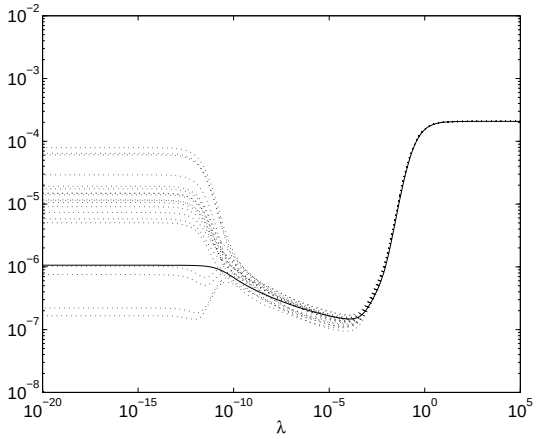

Figure 5.6: 20 replicates of $\overline{V}(\lambda)$ (dotted) and $E\overline{R}(\lambda) + \gamma\sigma^2$ for $\gamma = 0.1$, $\sigma = 0.001$ (solid)

Again for the same 20 replicates, Figure 5.6 shows the robust GCV functions $\overline{V}(\lambda)$ defined by (3.6) with $\gamma = 0.1$, together with the shifted robust risk $E\overline{R}(\lambda) + \gamma\sigma^2$ defined by (4.6). Clearly, the functions $\overline{V}(\lambda)$ track $E\overline{R}(\lambda) + \gamma\sigma^2$ in a neighbourhood of its minimizer, consistent with (4.8)

and Theorem 4.2. Comparing Figure 5.6 with Figure 5.3, note that the graphs of the functions $\overline{V}(\lambda)$ are modified sufficiently so that, for every replicate, the global minimizer is close to the minimizer of $E\overline{R}(\lambda)$.

To assess the accuracy and reliability of the methods, GCV and RGCV with $\gamma = 0.1$ were used to give parameter estimates $\hat{\lambda}_V$ and $\hat{\lambda}_{\overline{V}}$, 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) $I_R(\hat{\lambda}) = R(\hat{\lambda})/\min R(\lambda)$, (b) $I_{ER}(\hat{\lambda}) = ER(\hat{\lambda})/\min ER(\lambda)$,
(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 $R(\lambda) = n^{-1}\|K\boldsymbol{f}_\lambda - K\boldsymbol{f}_0\|^2$ and

$$R_1(\lambda) \;=\; \|K\boldsymbol{f}_\lambda - K\boldsymbol{f}_0\|_1^2 \;\equiv\; \sum_{i=1}^{n-2} n^{-2}(K\boldsymbol{f}_\lambda - K\boldsymbol{f}_0, \bar{\boldsymbol{\phi}}_i)^2 \bar{\lambda}_i^{-1}.$$

Very similar loss functions and inefficiencies were used in [19] for regularized solutions of (1.3).

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 $\boldsymbol{f}_\lambda$. This is because $R_1(\lambda)$ behaves like a squared discrete Sobolev seminorm of order 1 of the error $\boldsymbol{f}_\lambda - \boldsymbol{f}_0$. On the other hand, because the matrix K has the effect of smoothing high frequency error in the regularized solution $\boldsymbol{f}_\lambda$, the functions $R(\lambda)$ and $ER(\lambda)$ may be small even if the regularized solution is very noisy. A log-log plot of $\|K\sin(j\pi\boldsymbol{x})\|_1^2$ against j confirms this behaviour; $\|K\sin(j\pi\boldsymbol{x})\|_1^2$ behaves like cj^2, consistent with a squared Sobolev seminorm of order 1 of $\sin(j\pi\boldsymbol{x})$. In contrast, a log-log plot shows that $n^{-1}\|K\sin(j\pi\boldsymbol{x})\|^2$ behaves like cj^{-4}, as expected since K corresponds to integration twice. Therefore, of the four inefficiencies above, I_{R_1} is the best guide to the accuracy of the regularized solution.

For the GCV estimate $\hat{\lambda} = \hat{\lambda}_V$, histograms with bin width 0.5 for the four inefficiencies above are plotted in Figures 5.7 (a), (b), (c) and (d), respectively. If the inefficiency is greater than or equal to 50, it is included in the bin at 50. From the four histograms, for 60-80% of the replicates, the corresponding inefficiency satisfies $1 \le I \le 1.5$, 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 5.8 (a), (b), (c) and (d) show the inefficiency histograms of the RGCV (with $\gamma = 0.1$) estimate $\hat{\lambda} = \hat{\lambda}_{\overline{V}}$ for the same 100 replicates. Clearly, for $90 - 100\%$ of the replicates, the inefficiencies in (c) and (d) are less than 1.5, and there are very few replicates for which any inefficiency is large, so the method is reliable. Note that for the RGCV estimate, there is more spread in the inefficiencies I_R near 1 in Figure 5.8(a) than for the GCV estimate in Figure 5.7(a). This is because the RGCV method estimates the minimizer of the robust risk $E\overline{R}(\lambda)$ rather than the minimizer of the risk $ER(\lambda)$, and so it tends to give a slightly larger estimate $\hat{\lambda}_{\overline{V}}$ than the GCV estimate $\hat{\lambda}_V$ when the latter is good. However, note that the inefficiencies I_{R_1} for RGCV in Figure 5.8(c) are closer to 1 than those for GCV in Figure 5.7(c), indicating that RGCV has the more favourable accuracy.

To see the effect of the choice of γ, Figure 5.9 shows the proportion of the 100 replicates for which the RGCV estimate $\hat{\lambda} = \hat{\lambda}_{\overline{V}}$ gives inefficiency $I_{R_1}(\hat{\lambda}) > 1.5$ ($*$ symbol), $I_{R_1}(\hat{\lambda}) > 2$ ($+$ symbol) and $I_{R_1}(\hat{\lambda}) > 4$ ($\circ$ symbol). Clearly all these proportions tend to decrease significantly as γ decreases from 1 (the GCV case) until γ is about 0.1. This means there is a substantial

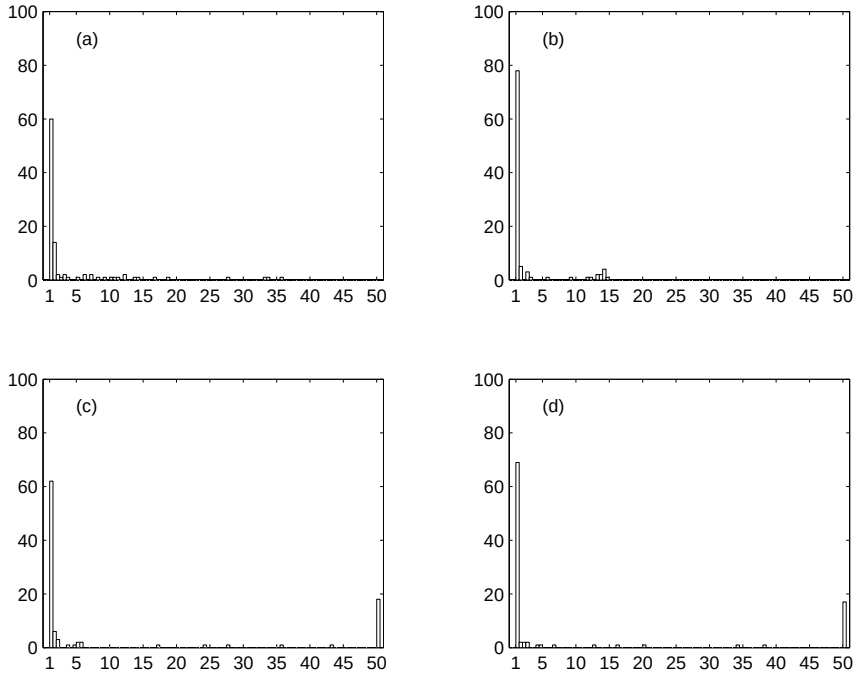

Figure 5.7: Histograms of inefficiencies (a) $I_R(\hat{\lambda})$, (b) $I_{ER}(\hat{\lambda})$, (c) $I_{R_1}(\hat{\lambda})$ and (d) $I_{ER_1}(\hat{\lambda})$ for GCV estimate $\hat{\lambda} = \hat{\lambda}_V$

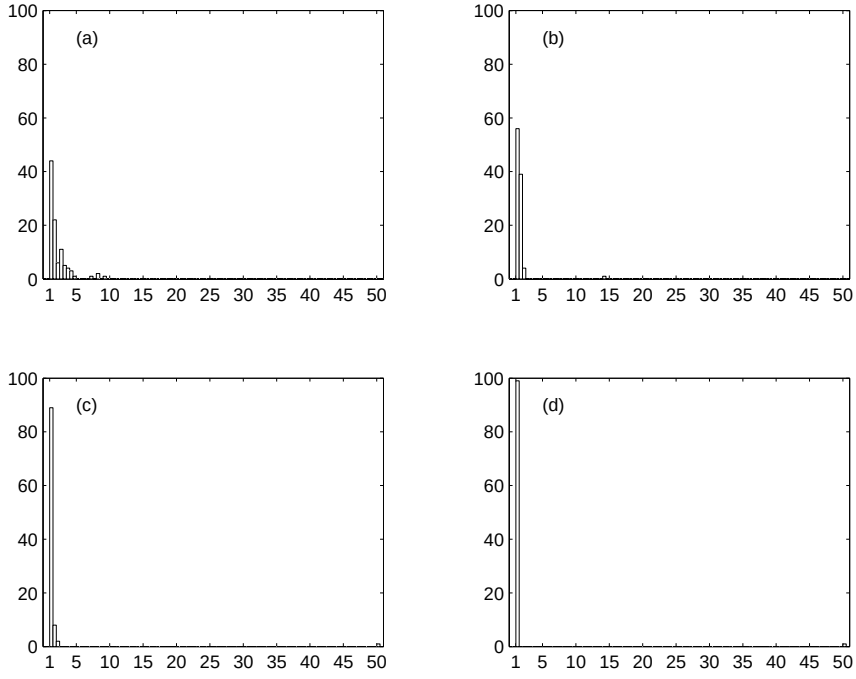

Figure 5.8: Histograms of inefficiencies (a) $I_R(\hat{\lambda})$, (b) $I_{ER}(\hat{\lambda})$, (c) $I_{R_1}(\hat{\lambda})$ and (d) $I_{ER_1}(\hat{\lambda})$ for RGCV ($\gamma = 0.1$) estimate $\hat{\lambda} = \hat{\lambda}_{\overline{V}}$

improvement in accuracy and reliability. It can be seen that if γ is chosen in $[0.1, 0.5]$, RGCV produces less than about half the number of poor regularized solutions as GCV. Note, however, that RGCV fails if γ is very close to 0 (the proportion with $I_{R_1} > 1.5$ increases significantly as $\gamma \to 0$). In this case, although the RGCV estimates are stable, generally they are too large giving inaccurate solutions.

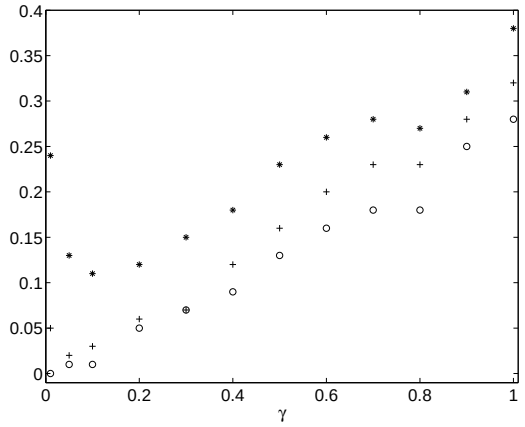

Figure 5.9: Proportion of RGCV replicates with $I_{R_1} > 1.5$ $(*)$, $I_{R_1} > 2$ $(+)$ and $I_{R_1} > 4$ $(\circ)$

References

[1] Cavalier L., Golubev, G.K., Picard D., Tsybakov, A.B.: Oracle inequalities for inverse problems. Ann. Statist. **30**, 843-874 (2002)

[2] Cook, R.D.: Detection of influential observation in linear regression, Technometrics **19**, 15-18 (1977)

[3] Cook, R.D., Weisberg, S.: Residuals and Influence in Regression, Chapman and Hall, New York, 1982

[4] Cox, D.D.: Approximation of method of regularization estimators. Ann. Statist. **16**, 694-712 (1988)

[5] Craven, P., Wahba, G.: Smoothing noisy data with spline functions. Numer. Math. **31**, 377-403 (1979)

[6] Efron, B.: Selection criteria for scatterplot smoothers. Ann. Statist. **29**, 470-504 (2001)

[7] Golub, G.H., Heath, M., Wahba, G.: Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics **21**, 215-223 (1979)

[8] Hall, P., Marron, J.S.: Local minima in cross-validation functions. J. Roy. Statist. Soc. Ser. B **53**, 245-252 (1991)

[9] Hall, P., Johnstone, I.: Empirical functionals and efficient smoothing parameter selection (with discussion). J. Roy. Statist. Soc. Ser. B **54**, 475-530 (1992)

[10] Hansen, P.C.: Regularization Tools: A Matlab package for analysis and solution of discrete ill-posed problems. Numerical Algorithms **6**, 1-35 (1992)

[11] Kimeldorf, G.S., Wahba, G.: Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. **33**, 82-95 (1971)

[12] Kohn, R., Ansley, C.F., Tharm, D.: The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. J. Amer. Statist. Assoc. **86**, 1042-1050 (1991)

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

[14] Kou, S.C.: From finite sample to asymptotics: a geometric bridge for selection criteria in spline regression. Ann. Statist. **32**, 2444-2468 (2004)

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

[16] Lin, Y., Brown, L.D.: Statistical properties of the method of regularization with periodic Gaussian reproducing kernel. Ann. Statist. **32**, 1723-1743 (2004)

[17] Lukas, M.A.: Convergence rates for regularized solutions. Math. Comp. **51**, 107-131 (1988)

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

[19] Lukas, M.A.: Comparisons of parameter choice methods for regularization with discrete noisy data. Inverse Problems **14**, 161-184 (1998)

[20] Lukas, M.A.: Strong robust generalized cross-validation for choosing the regularization parameter, in preparation.

[21] Nychka, D.W., Cox, D.D.: Convergence rates for regularized solutions of integral equations from discrete noisy data. Ann. Statist. **17**, 556-572 (1989)

[22] Robinson, T., Moyeed, R.: Making robust the cross-validation choice of smoothing parameter in spline smoothing regression. Commun. Statist.-Theory Meth. **18**, 523-539 (1989)

[23] Speckman, P.L., Sun D.: Asymptotic properties of smoothing parameter selection in spline smoothing. Preprint, 2001.

[24] Wahba, G.: Practical approximate solutions to linear operator equations when the data are noisy. SIAM J. Numer. Anal. **14**, 651-667 (1977)

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

[26] Wahba, G.: Spline Models for Observational Data, CBMS-NSF Regional Conference Series in Applied Mathematics 59, SIAM, Philadelphia, 1990

[27] Wahba, G., Wang, Y.D.: Behaviour near zero of the distribution of GCV smoothing parameter estimates. Statist. Prob. Lett. **25**, 105-111 (1995)