Robust GCV Choice of the Regularization Parameter for Correlated Data

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

Abstract. We consider Tikhonov regularization of linear inverse problems with discrete noisy data containing correlated errors. Generalized cross-validation (GCV) is a prominent parameter choice method, but it is known to perform poorly if the sample size n is small or if the errors are correlated, sometimes giving the extreme value 0. We explain why this can occur and show that the robust GCV methods perform better. In particular, it is shown that for any data set, there is a value of the robustness parameter below which the strong robust GCV method (R_1GCV) will not choose the value 0. We also show that, if the errors are correlated with a certain covariance model, then, for a range of values of the unknown correlation parameter, the "expected" R_1GCV estimate has a near optimal rate as $n \to \infty$. Numerical results for the problem of second derivative estimation are consistent with the theoretical results and show that R_1GCV gives reliable and accurate estimates.

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

1 Introduction

Consider the problem of estimating 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 ε_i are errors. In particular, we consider a linear ill-posed operator equation $Kf(x) = g(x)$, e.g. a first kind Fredholm integral equation, where the functionals are $L_i f = Kf(x_i)$, $i = 1, \ldots, n$. Another 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 $L_i \boldsymbol{f} = K \boldsymbol{f}_i$, where $\boldsymbol{f} \in \mathbb{R}^q$, $q \le n$, and K is the $n \times q$ model or design matrix.

In practical applications with observational data, it is appropriate to model the errors ϵ_i as random variables. Often it is assumed for simplicity that the errors are uncorrelated with zero mean (called white noise), but in actual fact the errors may have some correlation. This paper is mostly concerned with the latter situation. There are important applications in the geosciences, in particular, the estimation of the Earth's gravity field from satellite data [2].

To estimate the function f_0, we use Tikhonov regularization of the form [26]

$$\text{minimize} \quad n^{-1} \sum_{i=1}^{n} (L_i f - y_i)^2 \; + \; \lambda \|Pf\|_W^2 \tag{1.1}$$

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 $\|Pf\|_W^2 = \int (f^m(x))^2 dx$. For 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

$$\text{minimize} \quad n^{-1} \sum_{i=1}^{n} (K\boldsymbol{f}_i - y_i)^2 \; + \; \lambda \|M\boldsymbol{f}\|^2 \tag{1.2}$$

over $\boldsymbol{f} \in \mathbb{R}^q$, where $\|\cdot\|$ is the Euclidean norm and the matrix M is usually either I or a first or second order finite difference operator.

The accuracy of the regularized solution f_λ of (1.1) or (1.2) depends crucially on the choice of the regularization parameter λ. One of the most prominent methods for choosing the parameter is generalized cross-validation (GCV) due to Wahba [25]. GCV is known to have favorable asymptotic properties as $n \to \infty$ for uncorrelated data [25, 3, 12, 13].

However, GCV is not reliable when either n is small or the data are correlated. In these situations, it sometimes chooses 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 [23, 27]. For uncorrelated data, the robust GCV methods developed in [21, 16, 17] were shown to perform better than GCV for small n and have good asymptotic properties. In this paper, we investigate these methods for correlated data.

Let $A = A(\lambda)$ be the $n \times n$ influence matrix defined by $A\boldsymbol{y} = \boldsymbol{L}f_\lambda$, where $\boldsymbol{L}f = (L_1 f, \ldots, L_n f)^T$. Define $\mu_1(\lambda) = n^{-1}\mathrm{tr}\, A$, $\mu_2(\lambda) = n^{-1}\mathrm{tr}\, A^2$ and $\mu_{12}(\lambda) = -d\mu_1(\lambda)/d\lambda$. The GCV choice of λ is the minimizer of the GCV function

$$V(\lambda) \; = \; \frac{n^{-1}\|(I-A)\boldsymbol{y}\|^2}{[n^{-1}\mathrm{tr}(I-A)]^2} \; = \; \frac{n^{-1}\|(I-A)\boldsymbol{y}\|^2}{(1-\mu_1(\lambda))^2}. \tag{1.3}$$

If the covariance matrix of the errors is known, at least up to some parameterization, then the GCV function can be modified to include the covariance matrix, as described in [6, 20]. Similarly, in the context of wavelet thresholding, GCV can be extended to deal with correlated noise of a certain type [9]. However, in many situations the covariance matrix is unknown.

The robust GCV (RGCV) choice of λ is defined as the minimizer of the RGCV function

$$\overline{V}(\lambda) \; = \; \gamma V(\lambda) + (1-\gamma)F(\lambda) \; = \; (\gamma + (1-\gamma)\mu_2(\lambda))V(\lambda), \tag{1.4}$$

where $F(\lambda) = \mu_2(\lambda)V(\lambda)$ is an approximate average influence of all the data points on f_λ and where $\gamma \in (0, 1)$ is a robustness parameter. Another stabilized extension of GCV is the modified GCV method [4, 10, 24], which, under certain assumptions, is asymptotically equivalent to RGCV for uncorrelated data [17].

The strong robust GCV (R_1GCV) choice of λ is defined as the minimizer of the R_1GCV function

$$\overline{V}_1(\lambda) \; = \; \gamma V(\lambda) + (1-\gamma)F_1(\lambda) \; = \; (\gamma + (1-\gamma)\mu_{12}(\lambda))V(\lambda), \tag{1.5}$$

where $F_1(\lambda) = \mu_{12}(\lambda)V(\lambda)$ is an approximate total influence of all the data points measured in the W norm. In Section 2 we define spectral decompositions that can be used to compute $V(\lambda)$, $\overline{V}(\lambda)$ and $\overline{V}_1(\lambda)$.

In the case of uncorrelated data with small n, Efron [5, 11] used a geometric interpretation to explain the unstable behavior of GCV. The context in these papers was data smoothing, but the interpretation also applies in the regularization framework here. The same geometry is used in [18] to show that RGCV, with an appropriate value of γ, has much better stability than GCV.

For correlated data, it will be seen both in theory and in simulations that the behavior of GCV depends on the color of the noise. If the noise spectrum has greater power for lower frequencies, it is called red noise, while if the power is greater for higher frequencies, it is called blue noise. In the case of uncorrelated errors, i.e. white noise, the power spectrum is constant.

In Section 3, we use a sufficient condition to explain why GCV may choose the extreme value $\lambda = 0$ for small n or for strongly correlated data of red noise type. We also show how RGCV and R_1GCV can protect against this extreme choice. In particular, Theorem 3.4 shows that for all sufficiently small γ, the R_1GCV choice of λ is guaranteed to be positive.

In Section 4, we examine the asymptotic behavior of the R_1GCV method for Tikhonov regularization (1.1) of an ill-posed operator equation $Kf = g$ when the errors are correlated with a certain form of covariance matrix. Theorems 4.1 and 4.2 give the optimal rates for the prediction risk (mean square prediction error) and for a W norm risk. Theorem 4.3 shows that for white noise or red noise with a range of values of the correlation parameter, the (shifted) expected R_1GCV function tracks the strong robust risk in a neighbourhood of its minimizer, which has a near optimal decay rate. Therefore, no matter whether the errors are uncorrelated or correlated (something that may not be known in practice), R_1GCV has favorable asymptotic properties.

Section 5 describes numerical simulations for the discretized ill-posed problem of estimating the second derivative of a function $g(x)$ from noisy data $y_i = g(x_i) + \varepsilon_i$, $i = 1, \ldots, n$. This is the same example as in [16, 17]. The GCV, RGCV and R_1GCV estimates were computed for 200 replicates of the data, with both uncorrelated errors and correlated errors of different degrees of correlation. The numerical results are consistent with the theory. If the errors are uncorrelated or correlated of blue noise type, then all three criteria perform well, though GCV has a significant number of outliers. For correlated errors of red noise type, while GCV performs very poorly, both RGCV and R_1GCV perform well if the correlation is mild, and R_1GCV performs best by far if the correlation is strong.

2 Representation of robust GCV functions

Assume that the linear functionals $W \to \mathbb{R}$, $f \to L_i f$ are bounded and the null space $N(P)$ is finite dimensional with $N(\boldsymbol{L}) \cap N(P) = \{0\}$. Under these conditions it is well

known [26] that (1.1) has a unique solution, and the influence matrix has the form $A = Q(Q + n\lambda I)^{-1}$ if $P = I$ and

$$A = I - n\lambda B^T(B\Sigma B^T + n\lambda I)^{-1}B \tag{2.1}$$

if $P \neq I$, where Q and Σ are symmetric positive semidefinite $n \times n$ matrices and B is an $(n-m) \times n$ matrix satisfying $BB^T = I_{n-m}$.

As in [16, 17], we represent the GCV, RGCV and R_1GCV functions in terms of the following spectral decompositions. In the case where $P = I$, the matrix $n^{-1}Q$ 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 $\mathbb{R}^n$. For the problem of a first kind integral equation, these eigenvalues and eigenvectors are discretized approximations of the eigenvalues and L^2 normalized eigenfunctions of a certain integral operator [13].

In the case where $P \neq I$, there exists an $(n-m) \times (n-m)$ orthogonal matrix U such that $n^{-1}B\Sigma B^T = U\Lambda U^T$, where $\Lambda = \mathrm{diag}\{\bar{\lambda}_i, i = 1, \ldots, n-m\}$ and $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_{n-m} \geq 0$ (not all equal to 0). Let $W = B^TU$. Then $W^TW = I_{n-m}$, and from (2.1) we get $I - A = \lambda W(\Lambda + \lambda I)^{-1}W^T$. Let $\boldsymbol{w}_i$ be the ith column of W and define $\bar{\boldsymbol{\phi}}_i = \sqrt{n}\boldsymbol{w}_i$, $i = 1, \ldots, n-m$, so $n^{-1}(\bar{\boldsymbol{\phi}}_i, \bar{\boldsymbol{\phi}}_j) = \delta_{ij}$.

If $P \neq I$, the normalized residual sum of squares can be expressed as

$$n^{-1}\|(I - A)\boldsymbol{y}\|^2 = \lambda^2 \sum_{i=1}^{n-m} \hat{y}_i^2/(\bar{\lambda}_i + \lambda)^2, \tag{2.2}$$

where we denote $\hat{v}_i = n^{-1}(\boldsymbol{v}, \bar{\boldsymbol{\phi}}_i)$ for any vector $\boldsymbol{v}$. When $P = I$, the same equation (2.2) holds but with $m = 0$. Using the spectral decompositions above, the functions $\mu_1(\lambda)$, $\mu_2(\lambda)$ and $\mu_{12}(\lambda)$ can be expressed as

$$\mu_1(\lambda) \equiv n^{-1}\mathrm{tr}\, A = n^{-1}\left(m + \sum_{i=1}^{n-m} \bar{\lambda}_i/(\bar{\lambda}_i + \lambda)\right), \tag{2.3}$$

$$\mu_2(\lambda) \equiv n^{-1}\mathrm{tr}\, A^2 = n^{-1}\left(m + \sum_{i=1}^{n-m} [\bar{\lambda}_i/(\bar{\lambda}_i + \lambda)]^2\right) \quad \text{and} \tag{2.4}$$

$$\mu_{12}(\lambda) \equiv -\frac{d\mu_1(\lambda)}{d\lambda} = n^{-1} \sum_{i=1}^{n-m} \bar{\lambda}_i/(\bar{\lambda}_i + \lambda)^2 \tag{2.5}$$

if $P \neq I$, and the same expressions but with $m = 0$ if $P = I$. These expressions can be used to compute the GCV, RGCV and R_1GCV functions $V(\lambda)$, $\overline{V}(\lambda)$ and $\overline{V}_1(\lambda)$ defined in (1.3), (1.4) and (1.5), respectively.

Discrete regularization method

It is well known that for the fully discrete regularization problem (1.2), if $N(K) \cap N(M) = \{0\}$, there is a unique regularized solution $\boldsymbol{f}_\lambda = (K^TK + n\lambda M^TM)^{-1}K^T\boldsymbol{y}$, and the influence matrix is $A = K(K^TK + n\lambda M^TM)^{-1}K^T$.

In the case where $M = I_q$, the regularized solution and the GCV, RGCV and R_1GCV functions can be computed using the singular value decomposition (SVD) of K. In this case $n^{-1}\|(I-A)\boldsymbol{y}\|^2$ and the functions $\mu_1(\lambda)$, $\mu_2(\lambda)$ and $\mu_{12}(\lambda)$ are given by equations of the same form as (2.2) – (2.5) but with $m = 0$.

In the case where $M \neq I$ is a $p \times q$ matrix with $p \leq q \leq n$, it is known [7] 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). With appropriate definitions of $\bar{\lambda}_i$ and $\bar{\boldsymbol{\phi}}_i$ (see [16]), if $q = n$ (i.e. $M = M_{p\times n}$), then $n^{-1}\|(I-A)\boldsymbol{y}\|^2$, $\mu_1(\lambda)$, $\mu_2(\lambda)$ and $\mu_{12}(\lambda)$ can be expressed in the same form as in (2.2) – (2.5) but with $m = n - p$.

3 Extreme undersmoothing behavior

In this section, we investigate why GCV may choose the extreme value $\lambda = 0$ and how the RGCV and R_1GCV methods can protect against this. The results apply to both the regularization methods (1.1) and (1.2) with $M = I_q$ or $M = M_{p\times n}$. In these cases we have expressions of the form in (2.2) – (2.5), and we will write the results in the notation of these equations. The sums are from $i = 1$ to $i = n - m$ unless otherwise indicated.

3.1 GCV and robust GCV

The following result identifies important components in the behavior of the GCV and RGCV functions, including the effect of the parameter γ. For GCV, some parts of this result are derived in [23].

Lemma 3.1 *For all $\lambda > 0$, the derivative $\overline{V}'(\lambda)$ satisfies*

$$\overline{V}'(\lambda) \;=\; 2n^2\left(1 - S(\lambda) - (1-\gamma)[(1 - S(\lambda)(1-\mu_2(\lambda)) + n^{-1}T(\lambda)]\right)U(\lambda), \tag{3.1}$$

where

$$S(\lambda) \;=\; \frac{\sum[\hat{y}_i^2(\bar{\lambda}_i+\lambda)^{-1}](\bar{\lambda}_i+\lambda)^{-2}/\sum(\bar{\lambda}_i+\lambda)^{-2}}{\sum[\hat{y}_i^2(\bar{\lambda}_i+\lambda)^{-1}](\bar{\lambda}_i+\lambda)^{-1}/\sum(\bar{\lambda}_i+\lambda)^{-1}} \tag{3.2}$$

is the ratio of two different weighted averages of $\hat{y}_i^2(\bar{\lambda}_i+\lambda)^{-1}$, $i = 1, \ldots, n-m$, with $\hat{y}_i = n^{-1}(\boldsymbol{y}, \bar{\boldsymbol{\phi}}_i)$, and where

$$T(\lambda) \;=\; \frac{\sum(\bar{\lambda}_i+\lambda)^{-1}\sum\bar{\lambda}_i^2(\bar{\lambda}_i+\lambda)^{-3}}{\sum(\bar{\lambda}_i+\lambda)^{-2}} \tag{3.3}$$

and

$$U(\lambda) \;=\; \frac{\sum(\bar{\lambda}_i+\lambda)^{-2}\sum\hat{y}_i^2(\bar{\lambda}_i+\lambda)^{-2}}{(\sum(\bar{\lambda}_i+\lambda)^{-1})^3}. \tag{3.4}$$

If $\bar{\lambda}_i > 0$ for all $i = 1, \ldots, n-m$, then

$$\overline{V}'(0) \;=\; 2n^2[1 - S(0) - (1-\gamma)n^{-1}T(0)]U(0). \tag{3.5}$$

If $\bar{\lambda}_i > 0$ for $i \le \bar{n}$ and $\bar{\lambda}_i = 0$ for $i > \bar{n}$, where $\bar{n} < n - m$, then

$$\overline{V}'(0) \;=\; -2n^2(n-m-\bar{n})^{-3}\sum_{i=1}^{\bar{n}}\bar{\lambda}_i^{-1}\sum_{i=\bar{n}+1}^{n-m}\hat{y}_i^2, \tag{3.6}$$

so $\overline{V}'(0) < 0$ if the last sum is non-zero (which is almost certain in practice). As $\lambda \to \infty$,

$$\overline{V}'(\lambda) \;\sim\; \frac{2n^2(\gamma + (1-\gamma)m/n)(Y-1)}{(n-m)^3\lambda^2}\sum_{i=1}^{n-m}\bar{\lambda}_i\sum_{i=1}^{n-m}\hat{y}_i^2, \tag{3.7}$$

where

$$Y \;=\; 1 + \frac{\lim_{\lambda\to\infty}\lambda(1-S(\lambda))}{(n-m)^{-1}\sum\bar{\lambda}_i} \;=\; \frac{\sum\hat{y}_i^2\bar{\lambda}_i/\sum\bar{\lambda}_i}{(n-m)^{-1}\sum\hat{y}_i^2} \tag{3.8}$$

is the ratio of two different weighted averages of $\hat{y}_i^2$, $i = 1, \ldots, n-m$.

Proof. The expressions for $\overline{V}'(\lambda)$ in (3.1) – (3.4) follow from a straightforward calculation of the derivative of $\overline{V}(\lambda) = (\gamma + (1-\gamma)\mu_2(\lambda))V(\lambda)$, where $V(\lambda)$ is given by (1.3), and rearrangement. If $\bar{\lambda}_i > 0$ for all $i = 1, \ldots, n-m$, then $\mu_2(0) = 1$ and $\overline{V}'(0)$ can be found by direct substitution of $\lambda = 0$ in (3.1) – (3.4) giving (3.5). If $\bar{\lambda}_i > 0$ for $i \le \bar{n}$ and $\bar{\lambda}_i = 0$ for $i > \bar{n}$, where $\bar{n} < n - m$, then $S(\lambda)$ in (3.2) satisfies

$$S(\lambda) \sim 1 + \lambda(n-m-\bar{n})^{-1}\sum_{i=1}^{\bar{n}}\bar{\lambda}_i^{-1} \tag{3.9}$$

as $\lambda \to 0$. Also, as $\lambda \to 0$, we have $\mu_2(\lambda) \to n^{-1}(m+\bar{n})$,

$$T(\lambda) \sim \lambda\sum_{i=1}^{\bar{n}}\bar{\lambda}_i^{-1} \quad\text{and}\quad U(\lambda) \sim \lambda^{-1}(n-m-\bar{n})^{-2}\sum_{i=\bar{n}+1}^{n-m}\hat{y}_i^2. \tag{3.10}$$

Substituting these expressions into (3.1) and simplifying yields (3.6).

As $\lambda \to \infty$, we have $\mu_2(\lambda) \to m/n$,

$$S(\lambda) \sim 1 - \frac{1}{\lambda}\left(\frac{\sum\hat{y}_i^2\bar{\lambda}_i}{\sum\hat{y}_i^2} - (n-m)^{-1}\sum\bar{\lambda}_i\right), \tag{3.11}$$

$$T(\lambda) \sim \lambda^{-2}\sum\bar{\lambda}_i^2 \quad\text{and}\quad U(\lambda) \sim \lambda^{-1}(n-m)^{-2}\sum\hat{y}_i^2. \tag{3.12}$$

Substituting these expressions into (3.1) and rearranging yields (3.7) and (3.8). □

The next theorem shows why GCV may fail to choose a positive value of the regularization parameter. We will use the following lemma about weighted averages, which is proved in [15] and also follows easily from the discrete Chebyshev inequality [19, eq. (1.4), p. 240].

Lemma 3.2 *If $a_i > 0$, $b_i > 0$, $c_i > 0$, $i = 1, \ldots, N$, are such that the sequences $\{a_i\}$ and $\{b_i/c_i\}$ are non-constant and decreasing (i.e. $a_{i+1} \le a_i$ and $b_{i+1}/c_{i+1} \le b_i/c_i$), then*

$$\frac{\sum_{i=1}^N a_i b_i}{\sum_{i=1}^N b_i} > \frac{\sum_{i=1}^N a_i c_i}{\sum_{i=1}^N c_i}. \tag{3.13}$$

Theorem 3.1 *If $\bar{\lambda}_i > 0$ for all $i = 1, \ldots, n-m$ and the sequence $\hat{y}_i^2 \bar{\lambda}_i^{-1}$, $i = 1, \ldots, n-m$, is non-constant and decreasing, then $V'(\lambda) > 0$ for all $\lambda \geq 0$, so $V(\lambda)$ is minimized at $\lambda = 0$. If $\bar{\lambda}_i > 0$ for $i \leq \bar{n}$ and $\bar{\lambda}_i = 0$ for $i > \bar{n}$, where $\bar{n} < n - m$, and the sequences $\hat{y}_i^2 \bar{\lambda}_i^{-1}$, $i = 1, \ldots, \bar{n}$, and $\hat{y}_i^2$, $i = \bar{n}+1, \ldots, n$, are non-constant and decreasing, then $V'(\lambda) > 0$ for all $\lambda \geq \delta$ for some $\delta > 0$, which can be very small relative to the "optimal" parameter. In particular, if $\hat{y}_i^2 = ci^{-\rho}$ for $\rho > 0$, then $\delta = \rho^{-1}\bar{n}\bar{\lambda}_{\bar{n}}(1 + O(\bar{n}^{-1}))$, which, under the conditions of Theorem 4.2 and assuming $\bar{\lambda}_{\bar{n}} = O(n^{-r})$ for $r > 4/3$, satisfies $\delta/\lambda_W \to 0$ as $n \to \infty$, where λ_W is the optimal parameter for the W norm risk with uncorrelated errors.*

Proof. With $\gamma = 1$ in (3.5), we have that $\overline{V}'(0) = V'(0) > 0$ if and only if $S(0) < 1$. Clearly $S(0)$ is the ratio of two different weighted averages of the sequence $\hat{y}_i^2 \bar{\lambda}_i^{-1}$, $i = 1, \ldots, n-m$, with weights $\bar{\lambda}_i^{-2}$ and $\bar{\lambda}_i^{-1}$, $i = 1, \ldots, n-m$, respectively. By setting $a_i = \hat{y}_i^2 \bar{\lambda}_i^{-1}$, $b_i = \bar{\lambda}_i^{-1}$ and $c_i = \bar{\lambda}_i^{-2}$, Lemma 3.2 implies that $S(0) < 1$. (Intuitively, the weights $\bar{\lambda}_i^{-2}$ in the numerator of $S(0)$ put more weight on the smaller tail-end terms of the sequence $\hat{y}_i^2 \bar{\lambda}_i^{-1}$ compared to the weights $\bar{\lambda}^{-1}$ in the denominator.) Similarly, for $\lambda > 0$, with $\gamma = 1$ in (3.1), we have that $\overline{V}'(\lambda) = V'(\lambda) > 0$ if and only if $S(\lambda) < 1$. This follows from Lemma 3.2, by setting $a_i = \hat{y}_i^2(\bar{\lambda}_i + \lambda)^{-1}$, $b_i = (\bar{\lambda}_i + \lambda)^{-1}$ and $c_i = (\bar{\lambda}_i + \lambda)^{-2}$ (since clearly both $a_i = \hat{y}_i^2 \bar{\lambda}_i^{-1}[\bar{\lambda}_i/(\bar{\lambda}_i + \lambda)]$ and $b_i/c_i = \bar{\lambda}_i + \lambda$ are non-constant and decreasing). The second statement follows in the same way, since it is not hard to show that, if $\lambda \geq \delta$, where $\delta = \bar{\lambda}_{\bar{n}}(\hat{y}_{\bar{n}}^2/\hat{y}_{\bar{n}+1}^2 - 1)^{-1}$, then the sequence with terms $a_i = \hat{y}_i^2/(\bar{\lambda}_i + \lambda)$ for $i \leq \bar{n}$ and $a_i = \hat{y}_i^2/\lambda$ for $i > \bar{n}$ is non-constant and decreasing for all $i = 1, \ldots, n-m$. If $\hat{y}_i^2 = ci^{-\rho}$, then a binomial expansion yields $\delta = \rho^{-1}\bar{n}\bar{\lambda}_{\bar{n}}(1 + O(\bar{n}^{-1}))$. Since $\bar{\lambda}_{\bar{n}} = O(n^{-r})$ and $r > 3/2$, it follows from Theorem 4.2 that $\delta/\lambda_W = O(n^{1-r}n^{1/3}) \to 0$ as $n \to \infty$. □

It is clear from the proofs of Lemma 3.2 and Theorem 3.1 that if the sequence $\hat{y}_i^2 \bar{\lambda}_i^{-1}$, $i = 1, \ldots, n-m$, deviates only slightly from being decreasing, then it is still quite likely that $S(\lambda) < 1$ for all $\lambda \geq 0$, in which case $V(\lambda)$ is minimized at $\lambda = 0$.

Now we describe two situations in which the sequence $\hat{y}_i^2 \bar{\lambda}_i^{-1}$, $i = 1, \ldots, n-m$, has a decreasing trend, and so, from Theorem 3.1, it is quite likely that $V(\lambda)$ is minimized at $\lambda = 0$. Consider an operator equation $Kf_0(x) = g(x)$, where $g(x)$ is smooth, and let $\boldsymbol{g} = (g(x_1), \ldots, g(x_n))^T = \boldsymbol{L}f_0$. We will assume, as is usually the case, that the eigenvectors $\bar{\boldsymbol{\phi}}_i$ have mostly increasing frequency (measured say by the number of sign changes) with increasing i.

1. Uncorrelated small errors and small sample size

Suppose that n is small and the errors ε_i are realizations of uncorrelated (or slightly correlated) random variables with small standard deviation relative to $\|\boldsymbol{g}\|$. Since n is small, all the eigenvectors $\bar{\boldsymbol{\phi}}_i$, $i = 1, \ldots, n-m$, are of low frequency. Because $\boldsymbol{\varepsilon}$ is generally of high frequency, and since also the standard deviation is relatively small, then $|\hat{\epsilon}_i| << |\hat{g}_i|$ for all i. This implies that $\hat{y}_i^2 \bar{\lambda}_i^{-1} \approx \hat{g}_i^2 \bar{\lambda}_i^{-1}$, $i = 1, \ldots, n-m$. Now, from (2.14) in [17], using $\boldsymbol{L}f_0 = \boldsymbol{g}$, we have

$$\|P(f_0)_{\text{int}}\|_W^2 \;=\; \sum_{i=1}^{n-m} \hat{g}_i^2 \bar{\lambda}_i^{-1}, \tag{3.14}$$

where $(f_0)_{\rm int}$ is the solution of the generalized interpolation problem: minimize $\|Ph\|_W^2$ over $h \in W$ subject to $\boldsymbol{L}h = \boldsymbol{L}f_0$. It is known [14] that, under certain conditions, $\|P(f_0)_{\rm int}\|_W^2 \to \|Pf_0\|_W^2$ as $n \to \infty$, so the sum in (3.14) is bounded independent of n and hence the terms $\hat{g}_i^2\bar{\lambda}_i^{-1}$ have a decreasing trend. This is called a discrete Picard condition [8, 15]. Therefore, the sequence $\hat{y}_i^2\bar{\lambda}_i^{-1}$, $i = 1, \ldots, n-m$, has a decreasing trend and so it is quite likely that $V(\lambda)$ is minimized at $\lambda = 0$.

Note that, if instead n is large, then the sequence $\hat{y}_i^2\bar{\lambda}_i^{-1}$, $i = 1, \ldots, n-m$, does not have a decreasing trend for large i. This is because $\hat{g}_i \approx 0$ for such i (since $\bar{\boldsymbol{\phi}}_i$ is of high frequency and $g(x)$ is smooth), so $\hat{y}_i^2 \approx \hat{\epsilon}_i^2$, which does not approach 0 for large i since $\boldsymbol{\varepsilon}$ is not smooth. In fact, if the errors ε_i are random variables with mean 0, then $E\hat{y}_i^2 = \hat{g}_i^2 + E\hat{\epsilon}_i^2$, where E denotes expectation, and, if ε_i are uncorrelated with variance σ^2, then

$$E\hat{\epsilon}_i^2 \;=\; En^{-2}\sum_{j,k}\varepsilon_j\varepsilon_k(\bar{\boldsymbol{\phi}}_i)_j(\bar{\boldsymbol{\phi}}_i)_k \;=\; n^{-1}\sigma^2 n^{-1}\sum_j(\bar{\boldsymbol{\phi}}_i)_j^2 \;=\; n^{-1}\sigma^2,$$

so $E\hat{y}_i^2\bar{\lambda}_i^{-1}$ actually increases for large i. Also note that, if n is small but $n^{-1}\sigma^2$ is large relative to the smaller values of $\hat{g}_i^2$, then $E\hat{y}_i^2\bar{\lambda}_i^{-1}$ increases for i near $n-m$. The above observations indicate that, for either a larger sample size n or a larger error variance σ^2, GCV is less likely to choose the extreme value of 0.

2. Strongly correlated errors - red noise

Suppose that the errors ε_i are random variables with mean 0 and are correlated with covariance matrix $C = [E\varepsilon_i\varepsilon_j]$. Then

$$E\hat{\epsilon}_i^2 \;=\; n^{-2}\sum_{j,k}(\bar{\boldsymbol{\phi}}_i)_j C_{jk}(\bar{\boldsymbol{\phi}}_i)_k \;=\; n^{-2}\bar{\boldsymbol{\phi}}_i^T C\bar{\boldsymbol{\phi}}_i$$

and so

$$E\hat{y}_i^2\bar{\lambda}_i^{-1} = \hat{g}_i^2\bar{\lambda}_i^{-1} + n^{-2}\bar{\boldsymbol{\phi}}_i^T C\bar{\boldsymbol{\phi}}_i\bar{\lambda}_i^{-1}. \tag{3.15}$$

From above we can expect that $\hat{g}_i^2\bar{\lambda}_i^{-1}$ in (3.15) has a decreasing trend. Assume that $\varepsilon_i = \epsilon(x_i)$ for some noise process $\epsilon(x)$ with covariance function $E(\epsilon(s)\epsilon(t)) = Cov(s,t)$ that is at least continuously differentiable. Then the eigenvalues of $n^{-1}C$, which approximate those of Cov, decay quite quickly, and so the errors have significant correlation and are of red noise type. Since Cov is smooth and $n^{-1/2}\bar{\boldsymbol{\phi}}_i$, $i = 1, \ldots, n-m$, is an orthonormal (with respect to $(\cdot,\cdot)$) sequence of vectors of (mostly) increasing frequency, the sequence $n^{-2}\bar{\boldsymbol{\phi}}_i^T C\bar{\boldsymbol{\phi}}_i > 0$, $i = 1, \ldots, n-m$, also has a decaying behavior. If this decay is fast enough, $n^{-2}\bar{\boldsymbol{\phi}}_i^T C\bar{\boldsymbol{\phi}}_i\bar{\lambda}_i^{-1}$ also has a decreasing trend. Then, from (3.15), it is probable that $\hat{y}_i^2\bar{\lambda}_i^{-1}$ has a decreasing trend, and so it is quite likely that $V(\lambda)$ is minimized at $\lambda = 0$. Note that in this case, the conclusion is independent of the error variances and applies for both small and large sample size n. If, on the other hand, the sequence $n^{-2}\bar{\boldsymbol{\phi}}_i^T C\bar{\boldsymbol{\phi}}_i > 0$ has an increasing trend, i.e. the noise is blue, then, clearly, $\hat{y}_i^2\bar{\lambda}_i^{-1}$ cannot have a decreasing trend.

For the RGCV method, Lemma 3.1 gives the following result, which shows that, for any error behavior and sample size n, a smaller value of γ provides greater protection against the extreme choice of $\lambda = 0$.

Theorem 3.2 *For any $\lambda > 0$, if $S(\lambda) < 1$ (equivalently, if $V'(\lambda) > 0$), then $\overline{V}'(\lambda)$ decreases as γ decreases from 1. If $\bar{\lambda}_i > 0$ for all $i = 1, \ldots, n-m$, then $\overline{V}'(0)$ decreases as γ decreases from 1.*

Proof. The first part of the theorem follows from (3.1) since $\mu_2(\lambda) < 1$, $T(\lambda) > 0$ and $U(\lambda) > 0$. The second part follows from (3.5) since $T(0) > 0$ and $U(0) > 0$. □

In the special case where one or more of the $\bar{\lambda}_i$ equal 0, it is clear from (3.6) that $\overline{V}'(0)$ does not depend on γ and $\overline{V}'(0) < 0$. Consequently, the minimizers of $V(\lambda)$ and $\overline{V}(\lambda)$ must be positive.

Note that from (3.7) and (3.8), for any size n and any $\gamma \leq 1$ (including the GCV case of $\gamma = 1$), it is likely that $\overline{V}'(\lambda) > 0$ for all sufficiently large λ. This follows because the sequence $\hat{y}_i^2$, $i = 1, \ldots, n-m$, is close to non-constant and decreasing, and therefore, from Lemma 3.2 with $a_i = \hat{y}_i^2$, $b_i = \bar{\lambda}_i$ and $c_i = (n-m)^{-1}$, we have $Y > 1$ and so $\overline{V}'(\lambda) > 0$.

3.2 Strong robust GCV

The following result identifies important components in the behavior of the R_1GCV function.

Lemma 3.3 *For all $\lambda > 0$, the derivative $\overline{V}_1'(\lambda)$ satisfies*

$$\overline{V}_1'(\lambda) \;=\; 2n^2 \left(1 - S(\lambda) - (1-\gamma)[(1 - S(\lambda)(1-\mu_{12}(\lambda)) + n^{-1}T_1(\lambda)]\right) U(\lambda) \quad (3.16)$$

where $S(\lambda)$ and $U(\lambda)$ are defined in (3.2) and (3.4), and

$$T_1(\lambda) \;=\; \frac{\sum(\bar{\lambda}_i + \lambda)^{-1} \sum \bar{\lambda}_i(\bar{\lambda}_i + \lambda)^{-3}}{\sum(\bar{\lambda}_i + \lambda)^{-2}}. \quad (3.17)$$

If $\bar{\lambda}_i > 0$ for all $i = 1, \ldots, n-m$, then

$$\overline{V}_1'(0) \;=\; 2n^2 \left(1 - S(0) - (1-\gamma)[1 - S(0) + S(0)n^{-1}\sum \bar{\lambda}_i^{-1}]\right) U(0). \quad (3.18)$$

If $\bar{\lambda}_i > 0$ for $i \leq \bar{n}$ and $\bar{\lambda}_i = 0$ for $i > \bar{n}$, where $\bar{n} < n-m$, then

$$\overline{V}_1'(0) = \frac{-2n^2}{d^3} \left\{ \gamma \sum_{i=1}^{\bar{n}} \bar{\lambda}_i^{-1} + \frac{(1-\gamma)}{n} \left[\left(\sum_{i=1}^{\bar{n}} \bar{\lambda}_i^{-1} \right)^2 + d \sum_{i=1}^{\bar{n}} \bar{\lambda}_i^{-2} \right] \right\} \sum_{i=\bar{n}+1}^{n-m} \hat{y}_i^2, \quad (3.19)$$

where $d = n-m-\bar{n}$, so $\overline{V}_1'(0) < 0$ if the last sum is non-zero (which is almost certain in practice). As $\lambda \to \infty$,

$$\overline{V}_1'(\lambda) \sim \frac{2n^2\gamma(Y-1)}{(n-m)^3\lambda^2} \sum_{i=1}^{n-m} \bar{\lambda}_i \sum_{i=1}^{n-m} \hat{y}_i^2, \quad (3.20)$$

where Y is defined in (3.8).

Proof. The expressions for $\overline{V}_1(\lambda)$ in (3.16) – (3.17) follow from a straightforward differentiation of $\overline{V}_1(\lambda) = (\gamma + (1-\gamma)\mu_{12}(\lambda))V(\lambda)$, where $V(\lambda)$ is given in (1.3), and rearrangement. If $\bar{\lambda}_i > 0$ for all $i = 1, \ldots, n-m$, then $\overline{V}_1'(0)$ can be found by direct substitution of $\lambda = 0$ in (3.16) giving (3.18), since $\mu_{12}(0) = n^{-1}\sum \bar{\lambda}_i^{-1} = n^{-1}T_1(0)$. If $\bar{\lambda}_i > 0$ for $i \le \bar{n}$ and $\bar{\lambda}_i = 0$ for $i > \bar{n}$, where $\bar{n} < n-m$, then

$$\mu_{12}(\lambda) \to n^{-1}\sum_{i=1}^{\bar{n}} \bar{\lambda}_i^{-1} \quad \text{and} \quad T_1(\lambda) \sim \lambda \sum_{i=1}^{\bar{n}} \bar{\lambda}_i^{-2}$$

as $\lambda \to 0$. Substituting these expressions and those for $S(\lambda)$ in (3.9) and $U(\lambda)$ in (3.10) into (3.16) and simplifying yields (3.19).

As $\lambda \to \infty$, clearly $\mu_{12}(\lambda) \to 0$ and $T_1(\lambda) \sim \lambda^{-2}\sum_{i=1}^{n-m} \bar{\lambda}_i$. Substituting these expressions and those for $S(\lambda)$ in (3.11) and $U(\lambda)$ in (3.12) into (3.16) and rearranging yields (3.20). □

From Lemma 3.3 we get the following result for R_1GCV, which (like Theorem 3.2) shows that a smaller value of γ provides greater protection against the extreme choice of $\lambda = 0$.

Theorem 3.3 *For any $\lambda > 0$, if $(1 - S(\lambda)(1-\mu_{12}(\lambda)) + n^{-1}T_1(\lambda) > 0$, then $\overline{V}_1'(\lambda)$ decreases as γ decreases from 1. If $\bar{\lambda}_i > 0$ for $i = 1, \ldots, n-m$ and $1 + S(0)(-1 + n^{-1}\sum \bar{\lambda}_i^{-1}) > 0$ (which holds if $\bar{\lambda}_i$ decays sufficiently quickly), then $\overline{V}_1'(0)$ decreases as γ decreases from 1.*

Proof. The first part of the theorem follows from (3.16) and the second part follows from (3.18). □

Further to Theorem 3.3, the following result shows that, whatever the error behavior or value of n, we can ensure that $\overline{V}_1'(0) < 0$ by taking γ sufficiently small. Therefore, for all γ sufficiently small, the R_1GCV method is guaranteed to choose a positive regularization parameter. This is not true for the RGCV method, since in (3.5) the value of $1 - S(0)$ may be larger than $n^{-1}T(0)$ (which satisfies $n^{-1}T(0) \le n^{-1}(n-m)$ by the Cauchy-Schwartz inequality).

Theorem 3.4 *If $\bar{\lambda}_i = 0$ for some i, then $\overline{V}_1'(0) < 0$ for all $0 < \gamma \le 1$. If $\bar{\lambda}_i > 0$ for all i and if $S(0) > 1$, then $\overline{V}_1'(0) < 0$ for all $0 < \gamma \le 1$. If $\bar{\lambda}_i > 0$ for all i and $S(0) \le 1$, then $\overline{V}_1'(0) < 0$ for all*

$$\gamma \;<\; \frac{S(0)n^{-1}\sum_{i=1}^{n-m} \bar{\lambda}_i^{-1}}{1 - S(0) + S(0)n^{-1}\sum_{i=1}^{n-m} \bar{\lambda}_i^{-1}}.$$

Proof. The first part follows from (3.19). The second part follows from (3.18) written as

$$\overline{V}_1'(0) \;=\; -2n^2\left(\gamma(S(0)-1) + (1-\gamma)S(0)n^{-1}\sum \bar{\lambda}_i^{-1}\right)U(0).$$

The third part also follows from (3.18) by solving $\overline{V}_1'(0) < 0$ for γ. □

4 Asymptotic analysis

The framework for our asymptotic analysis is the same as that in [13, 16, 17]. 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.

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 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 $\tau_i \to \infty$.

For the "smoothness" class of f_0, we use the family of Hilbert spaces W_β 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 [13] that $W_1 = W$ with equivalent norms. Under certain conditions, the spaces W_β can be identified as fractional Sobolev spaces in which the smoothness increases with β [13].

We now state the main assumptions in this section. Assumption 4.1 specifies the error behavior, while Assumptions 4.2 – 4.5 are the same as those in [13, 16, 17] for the asymptotic analysis of GCV, RGCV and R_1GCV in the case of uncorrelated errors. 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 $a_n \lesssim b_n$ if there exists a positive constant c such that $a_n \le c b_n$.

Assumption 4.1 The errors ε_i are random variables with mean $E\varepsilon_i = 0$ and covariance matrix $C = [E\varepsilon_i \varepsilon_j]$ of the form

$$C = \sigma^2 (kn^{-1}Q)^t \text{ if } P = I \quad \text{and} \quad C = \sigma^2 (I_n - W(I_{n-m} - (k\Lambda)^t) W^T) \text{ if } P \ne I,$$

for some constants $k > 0$ and t, where Q, W and Λ are defined in Section 2. Clearly, when $t = 0$, the errors are uncorrelated. As $|t|$ increases from 0, the errors become increasingly correlated, with blue noise for $t < 0$ and red noise for $t > 0$.

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 $q(x,x) \leq \bar{q}$ 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 $f, v \in W$

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

Assumption 4.6 There is a sequence $\alpha_n' \to 0$ such that, for any p satisfying $1/r < p < 2 - 1/r$,

$$n^{-1} \sum_{i=1}^{n-m} \bar{\lambda}_i^p (\bar{\lambda}_i + \lambda)^{-2} \;\approx\; n^{-1} D(\lambda; p - 2 - 1/r, -2),$$

uniformly in $\lambda \in [\alpha_n', \infty)$ as $n \to \infty$, where $D(\lambda; a, b) \equiv \lambda^a$, if $\lambda \leq 1$, and $D(\lambda; a, b) \equiv \lambda^b$, if $\lambda > 1$. This is similar to the corresponding assumption made in [15].

Assumption 4.7 For each $t < 1/r$, as $n \to \infty$,

$$\nu(n) \equiv n^{-1} \left(m + \sum_{i=1}^{n-m} \bar{\lambda}_i^t \right) \approx n^{-1} \sum_{i=1}^{n-m} i^{-rt} \approx n^{-rt},$$

where the last estimate comes from an integral comparison.

The asymptotic analysis of the R_1GCV method depends crucially on the asymptotic behavior of the functions $\mu_1(\lambda)$, $\mu_2(\lambda)$ and $\mu_{12}(\lambda)$ defined in (2.3) – (2.5). The following estimates were derived in Theorems 4.1 and 4.2 of [13]. 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

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

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

$$\mu_{12}(\lambda) \;\approx\; n^{-1} D(\lambda; -(1+1/r), -2), \tag{4.3}$$

uniformly in $\lambda \in [\alpha_n, \infty)$. Note that the asymptotic estimate of $\mu_{12}(\lambda)$ in (4.3) is a particular case of Assumption 4.6.

4.1 Optimal parameter estimates

First we derive an estimate of the prediction risk $ER(\lambda) = En^{-1}\|\boldsymbol{L}f_\lambda - \boldsymbol{L}f_0\|^2$ and its minimizer in the case of correlated data.

Theorem 4.1 *Suppose that Assumptions 4.1 – 4.6 hold,* $-2 < t < 1/r$, $f_0 \in W_2$ *and* $\alpha_n \to 0$ *such that* $d_n^2 \alpha_n^{-(s+1)} \to 0$. *Then*

$$ER(\lambda) \;\approx\; \min\{1, \lambda^2\} + k^t \sigma^2 n^{-1} D(\lambda; t - 1/r, -2), \tag{4.4}$$

uniformly in $\lambda \in [\max\{\alpha_n, \alpha_n'\}, \infty)$. *Define* $\lambda^* = (\sigma^2 n^{-1})^{r/((2-t)r+1)}$ *and assume that* $\lambda^* \geq \max\{\alpha_n, \alpha_n'\}$. *Then the minimum of* $ER(\lambda)$ *for* $\lambda \geq \max\{\alpha_n, \alpha_n'\}$ *occurs at* $\lambda_R \approx \lambda^*$ *and* $ER(\lambda_R) \approx (\sigma^2 n^{-1})^{2r/((2-t)r+1)}$ *as* $n \to \infty$.

Proof. Since the errors ε_i have mean 0, we have $ER(\lambda) = b^2(\lambda) + v(\lambda)$, where $b^2(\lambda) = n^{-1}\|E\boldsymbol{L}f_\lambda - \boldsymbol{L}f_0\|^2$ is the squared bias and $v(\lambda) = En^{-1}\|A\boldsymbol{\varepsilon}\|^2$ is the variance. It is known (see Theorem 4.5 in [13]) that, since $f_0 \in W_2$, the squared bias satisfies $b^2(\lambda) \approx \min\{1, \lambda^2\}$, uniformly in $\lambda \in [\alpha_n, \infty)$. The variance satisfies

$$v(\lambda) = n^{-1} \sum_{i,j=1}^{n} C_{ij} A_{ij}^2 = n^{-1} \mathrm{tr} CA^2$$

and, from Assumptions 4.1 and 4.6, we obtain

$$v(\lambda) = \sigma^2 n^{-1} \left(m + k^t \sum_{i=1}^{n-m} \bar{\lambda}_i^{2+t} (\bar{\lambda}_i + \lambda)^{-2} \right) \;\approx\; k^t \sigma^2 n^{-1} D(\lambda; t - 1/r, -2),$$

uniformly in $\lambda \in [\max\{\alpha_n, \alpha_n'\}, \infty)$. The estimate (4.4) of $ER(\lambda)$ follows. Let $Y(\lambda)$ denote the right hand side of (4.4). Clearly, the minimum of $Y(\lambda)$ for $\lambda \geq \max\{\alpha_n, \alpha_n'\}$ occurs at $\lambda \approx \lambda^*$, and $\min Y(\lambda) \approx Y(\lambda^*)$. Also

$$\min Y(\lambda) \leq Y(\lambda_R) \approx ER(\lambda_R) \leq ER(\lambda^*) \approx Y(\lambda^*)$$

so $ER(\lambda_R) \approx Y(\lambda_R) \approx Y(\lambda^*) \approx (\sigma^2 n^{-1})^{2r/((2-t)r+1)}$. This implies that $\lambda_R \to 0$ as $n \to \infty$ since $\lambda_R^2 \leq Y(\lambda_R)$. Then, by substituting $\lambda_R = c_n \lambda^*$ into the relation $Y(\lambda_R) \approx Y(\lambda^*)$, we get $(c_n^2 + c_n^{t-1/r})\lambda^{*2} \approx \lambda^{*2}$, which implies that $c_n \approx 1$, and hence $\lambda_R \approx \lambda^*$. □

Note that for $t = 0$ (i.e. uncorrelated errors), the estimate $\lambda_R \approx (\sigma^2 n^{-1})^{r/(2r+1)}$ from Theorem 4.1 is the same as in Corollary 4.1 in [13]. Clearly, as t increases, the parameters λ^* and λ_R decay more quickly as $n \to \infty$.

Because the prediction risk only involves deviations in $\boldsymbol{L}f_\lambda$, it is a rather weak measure of the accuracy of f_λ. Consequently, we will also consider the stronger W norm risk defined as

$$ER_W(\lambda) = ER(\lambda) + E\|Pf_\lambda - Pf_0\|_W^2.$$

For example, the last term could be $E\|f_\lambda'' - f_0''\|_{L^2}^2$.

Theorem 4.2 *Suppose that Assumptions 4.1 – 4.6 hold,* $-2 < t < 1/r$, $f_0 \in W_3$ *and* $\alpha_n \to 0$ *such that* $d_n^2 \alpha_n^{-(s+1+1/r)} \to 0$. *Then*

$$ER_W(\lambda) \;\approx\; \min\{1, \lambda^2\} + k^t \sigma^2 n^{-1} D(\lambda; t - 1 - 1/r, -2), \tag{4.5}$$

uniformly in $\lambda \in [\max\{\alpha_n, \alpha'_n\}, \infty)$. *Define* $\lambda_W^* = (\sigma^2 n^{-1})^{r/((3-t)r+1)}$ *and assume that* $\lambda_W^* \geq \max\{\alpha_n, \alpha'_n\}$. *Then the minimum of* $ER_W(\lambda)$ *for* $\lambda \geq \max\{\alpha_n, \alpha'_n\}$ *occurs at* $\lambda_W \approx \lambda_W^*$ *and* $ER_W(\lambda_W) \approx (\sigma^2 n^{-1})^{2r/((3-t)r+1)}$ *as* $n \to \infty$.

Proof. Since the errors ε_i have mean 0, we have $ER_W = b^2 + v + b_1^2 + v_1$, where b^2 and v are defined in the proof of Theorem 4.1, and $b_1^2 = \|EPf_\lambda - Pf_0\|_W^2$ and $v_1 = E\|Pf_\lambda - EPf_\lambda\|_W^2$. It is known (see Proposition 3.1 in [13] with $\rho = 1$) that, since $f_0 \in W_3$, the squared bias b_1^2 satisfies $b_1^2(\lambda) \lesssim \min\{1, \lambda^2\}$, uniformly in $\lambda \in [\alpha_n, \infty)$. From equation (A.8) in [13], the variance v_1 satisfies $v_1 = \mathrm{tr} CF^T \Sigma F$, where $F = B^T(B\Sigma B^T + n\lambda I)^{-1}B$. Using the spectral decomposition of $B\Sigma B^T$ in Section 2 with Assumptions 4.1 and 4.6, we obtain

$$v_1(\lambda) = \sigma^2 n^{-1} \sum_{i=1}^{n-m} \bar{\lambda}_i^{1+t} (\bar{\lambda}_i + \lambda)^{-2} \approx \sigma^2 n^{-1} D(\lambda; t - 1 - 1/r, -2),$$

uniformly in $\lambda \in [\alpha'_n, \infty)$. Combining these estimates of $b_1^2(\lambda)$ and $v_1(\lambda)$ with the estimates of $b^2(\lambda)$ and $v(\lambda)$ in the proof of Theorem 4.1, and using $n^{-1}\lambda^{t-1-1/r} \approx v_1 \leq v + v_1 \lesssim 2n^{-1}\lambda^{t-1-1/r}$ for any $\lambda \leq 1$, yields the estimate (4.5) of $ER_W(\lambda)$. The remaining parts of the theorem follow in the same way as in the proof of Theorem 4.1. □

Note that for $t = 0$ (i.e. uncorrelated errors), the estimate $\lambda_W \approx (\sigma^2 n^{-1})^{r/(3r+1)}$ from Theorem 4.2 is the same as the optimal rate for the expected squared W norm error given in Corollary 3.1 in [13]. For the W norm risk (as for the prediction risk), as t increases, the optimal parameter λ_W decays faster to 0 as $n \to \infty$.

4.2 Asymptotic behavior of R_1GCV

For GCV with uncorrelated data, it is well known [26] that, as $n \to \infty$, the function $EV(\lambda) - \sigma^2$ tracks the prediction risk $ER(\lambda)$ in a neighbourhood of the optimal parameter for the risk. This is not true for correlated data.

For RGCV with uncorrelated data, as $n \to \infty$, the function $E\overline{V}(\lambda) - \gamma\sigma^2$ tracks the robust prediction risk $E\overline{R}(\lambda) \equiv \gamma ER(\lambda) + (1-\gamma)v(\lambda)$ in a neighbourhood of its minimizer, where $v(\lambda) = En^{-1}\|\boldsymbol{L}f_\lambda - E\boldsymbol{L}f_\lambda\|^2$ is the variance [16]. A similar result holds for R_1GCV with uncorrelated data: as $n \to \infty$, the function $E\overline{V}_1(\lambda) - \gamma\sigma^2$ tracks the strong robust risk $E\overline{R}_1(\lambda) \equiv \gamma ER(\lambda) + (1-\gamma)v_1(\lambda)$ in a neighbourhood of its minimizer, where $v_1(\lambda) = E\|Pf_\lambda - EPf_\lambda\|_W^2 = \sigma^2\mu_{12}(\lambda)$ is the variance [17]. We will show that for R_1GCV, an extension of this result also holds for correlated data.

Define the strong robust risk for correlated data, with covariance defined in Assumption 4.1, as

$$E\overline{R}_1(\lambda) = \gamma ER(\lambda) + (1-\gamma)\sigma^2\nu(n)\mu_{12}(\lambda), \tag{4.6}$$

where $\nu(n)$ is defined in Assumption 4.7. Note that this agrees with $E\overline{R}_1(\lambda)$ in the uncorrelated case, since $\nu(n) = 1$ when $t = 0$.

Theorem 4.3 *Suppose that Assumptions 4.1 – 4.7 hold, $f_0 \in W_3$ and $\alpha_n \to 0$ such that $d_n^2 \alpha_n^{-(s+1+1/r)} \to 0$. Also assume that $-2 < t < 1-(1-1/r)^{1/2}$. Define $\lambda^*_{\overline{R}_1} = n^{-r(1+rt)/(3r+1)}$ and assume that $\lambda^*_{\overline{R}_1} \geq \max\{\alpha_n, \alpha'_n\}$. Then the minimum of $E\overline{R}_1(\lambda)$ for $\lambda \geq \max\{\alpha_n, \alpha'_n\}$ occurs at $\lambda_{\overline{R}_1} \approx \lambda^*_{\overline{R}_1}$ as $n \to \infty$, and*

$$E\overline{V}_1(\lambda) - \gamma\sigma^2\nu(n) = E\overline{R}_1(\lambda)(1+o(1)) \tag{4.7}$$

for λ in a neighbourhood of $\lambda_{\overline{R}_1}$. If $0 < t < 1-(1-1/r)^{1/2}$, then $\lambda^* < \lambda_{\overline{R}_1} < \lambda^*_W$ for sufficiently large n; if $t = 0$, then $\lambda_{\overline{R}_1} \approx \lambda^*_W$; while if $-2 < t < 0$, then $\lambda_{\overline{R}_1} > \lambda^*_W$, where λ^* and λ^*_W are defined in Theorems 4.1 and 4.2, respectively.

Proof. From (4.6) and Assumption 4.1, as in the proof of Theorem 4.1, we have

$$E\overline{R}_1(\lambda) = \gamma b^2 + \gamma\sigma^2\mu_{2+t,2} + (1-\gamma)\sigma^2\nu(n)\mu_{12},$$

where $\mu_{2+t,2}$ is defined by

$$\mu_{p+t,q} = n^{-1}\left(m + k^t \sum_{i=1}^{n-m} \bar{\lambda}_i^{p+t}(\bar{\lambda}_i + \lambda)^{-q}\right). \tag{4.8}$$

Using $b^2 = \min\{1, \lambda^2\}$ and Assumptions 4.6 and 4.7, we obtain

$$E\overline{R}_1(\lambda) \approx Z(\lambda) \equiv \gamma\lambda^2 + \gamma\sigma^2 n^{-1}\lambda^{t-1/r} + (1-\gamma)\sigma^2 n^{-1-rt}\lambda^{-1-1/r} \tag{4.9}$$

for $\max\{\alpha_n, \alpha'_n\} \leq \lambda \leq 1$. Let λ_Z be the minimizer of $Z(\lambda)$. The last term of (4.9) is $\gtrsim$ the second last term if and only if $\lambda \lesssim n^{-rt/(1+t)}$ as $n \to \infty$. The minimizer λ_Z satisfies this condition, because, if it did not, it would be defined by $\lambda^2 \approx n^{-1}\lambda^{t-1/r}$, which leads to a contradiction since $t < 1-(1-1/r)^{1/2}$ (and so $rt^2 - 2rt + 1 > 0$). Hence, λ_Z is defined by $\lambda^2 \approx n^{-1-rt}\lambda^{-1-1/r}$, which gives

$$\lambda_Z \approx n^{-r(1+rt)/(3r+1)} = \lambda^*_{\overline{R}_1}.$$

This is consistent with the above condition since $t < 1-(1-1/r)^{1/2}$. Then, from (4.9), $E\overline{R}_1(\lambda_Z) \to 0$ as $n \to \infty$, and so $\lambda_{\overline{R}_1} \to 0$ (since, otherwise, $E\overline{R}_1(\lambda) \gtrsim 1$ for all λ). Therefore, we get $E\overline{R}_1(\lambda_{\overline{R}_1}) \approx Z(\lambda_{\overline{R}_1}) \approx Z(\lambda_Z)$ and $\lambda_{\overline{R}_1} \approx \lambda_Z$ by using (4.9) and the same argument as in the proof of Theorem 4.1.

For the numerator in (1.3), Assumption 4.1 gives

$$n^{-1}E\|(I-A)\boldsymbol{y}\|^2 = n^{-1}\mathrm{tr}C(I-A)^2 = \sigma^2(\nu(n) - 2\mu_{1+t,1} + \mu_{2+t,2}), \tag{4.10}$$

where $\mu_{1+t,1}$ and $\mu_{2+t,2}$ are defined by (4.8). Using (4.6), (1.3), (1.5) and (4.10), and rearranging, we obtain

$$\frac{E\overline{R}_1(\lambda) + \gamma\sigma^2\nu(n) - E\overline{V}_1(\lambda)}{E\overline{R}_1(\lambda)} = -\frac{\mu_1(2-\mu_1) + ((1-\gamma)/\gamma)\mu_{12}}{(1-\mu_1)^2} +$$

$$\frac{\sigma^2[2\gamma\mu_{1+t,1} - \gamma\mu_1(2-\mu_1)\nu(n) + 2(1-\gamma)\mu_{12}\mu_{1+t,1} + ((1-\gamma)^2/\gamma)\mu_{12}^2\nu(n)]}{E\overline{R}_1(\lambda)(1-\mu_1)^2}.$$

From (4.1), clearly $\mu_1(\lambda) \to 0$ as $n \to \infty$ for λ in a neighbourhood of $\lambda_{\overline{R}_1}$. By writing $\mu_{1+t,1}$ in terms of $\mu_{1+t,2}$ and $\mu_{2+t,2}$, and using Assumption 4.6, we find the estimate $\mu_{1+t,1} \approx n^{-1}\lambda^{t-1/r}$. Then, using Assumption 4.7 and the estimates of μ_1 and μ_{12} in (4.1) and (4.3), we obtain the bound

$$\frac{|E\overline{R}_1(\lambda) + \gamma\sigma^2\nu(n) - E\overline{V}_1(\lambda)|}{E\overline{R}_1(\lambda)} \lesssim 2\mu_1 + ((1-\gamma)/\gamma)\mu_{12} +$$

$$\frac{2\gamma\mu_{1+t,1} + 2\gamma\mu_1\nu(n) + 2(1-\gamma)\mu_{12}\mu_{1+t,1} + ((1-\gamma)^2/\gamma)\mu_{12}^2\nu(n)}{(1-\gamma)\mu_{12}\nu(n)}$$

$$\lesssim n^{-1}\lambda^{-1/r} + n^{-1}\lambda^{-1-1/r} + n^{rt}\lambda^{1+t} + \lambda + n^{-1+rt}\lambda^{t-1/r} + n^{-1}\lambda^{-1-1/r}.$$

Substituting $\lambda = \lambda_{\overline{R}_1}$, it is not hard to verify that all the terms in this bound approach 0 as $n \to \infty$ provided t satisfies $t < 2/(r+1)$, $rt^2 - 2rt + 1 > 0$ and $rt^2 - (3r+1)t + 3 > 0$, and all these inequalities hold if $t < 1 - (1-1/r)^{1/2}$. This shows (4.7). The last statement follows by comparing the estimate of $\lambda_{\overline{R}_1}$ with λ^* and λ_W^*. □

Since $E\overline{V}_1(\lambda)$ and $E\overline{V}_1(\lambda) - \gamma\sigma^2\nu(n)$ have the same minimizer, Theorem 4.3 indicates that there is an "expected" R_1GCV estimate $\lambda_{\overline{V}_1}$ that behaves like $\lambda_{\overline{R}_1}$ for large n. Moreover, in the problematic case of red noise ($t > 0$) (see Section 3.1), for a range of t values, $\lambda_{\overline{V}_1}$ has near optimal performance for large n. On the other hand, for blue noise ($t < 0$), $\lambda_{\overline{V}_1}$ is oversmoothing for large n, though not by much if $|t|$ is small, since, when $t = 0$, $\lambda_{\overline{V}_1}$ behaves like $\lambda_{\overline{R}_1} \approx \lambda_W^*$. Note that $1 - (1-1/r)^{1/2}$ decreases (from nearly 1) as $r > 1$ increases, so the range of allowable t values for red noise becomes smaller with greater degree of ill-posedness.

5 Numerical results

We consider the ill-posed problem and method in [16, 17] of estimating the second derivative function $f(x) = g''(x)$, $0 \le x \le 1$, from discrete noisy data. Assuming $g(0) = g(1) = 0$, the second derivative satisfies the first kind Fredholm integral equation $\int_0^1 k(x,t)f(t)\,dt = g(x)$, where $k(x,t) = x(t-1)$ if $x < t$ and $k(x,t) = t(x-1)$ if $x \ge t$. After discretization using the trapezoidal rule and uniform collocation points $x_i = (i-1)/(n-1)$, $i = 1, \ldots, n$, the equation becomes $K\boldsymbol{f} = \boldsymbol{g}$ for an $n \times n$ matrix K. Then Tikhonov regularization of the form (1.2) is applied with an $n \times n$ first order difference matrix M.

We take $g(x) = (x^3 - x)/6$, so the solution is $f_0(x) = x$, and generate data $y_i = (K\boldsymbol{f}_0)_i + \varepsilon_i$, $i = 1, \ldots, n$, with pseudo-random normal errors ε_i with mean 0. We assume the errors are either uncorrelated with equal variance σ^2 as in [16, 17] or they satisfy the first order autoregressive (AR(1)) correlation model with covariance matrix defined by $Cov_{ij} = E(\varepsilon_i\varepsilon_j) = \sigma^2\omega^{|i-j|}$ for $-1 < \omega < 1$, $\omega \ne 0$. Clearly, if $\omega < 0$, adjacent errors are negatively correlated, and if $\omega > 0$, they are positively correlated. In the latter case, the errors are red noise and the model is a discrete version of the Ornstein-Uhlenbeck process [22]. This correlation model was used for nonparametric

regression in [6, 20]. Figure 1 shows the function $g(x)$ and correlated data y_i from the AR(1) model with $n = 101$, $\sigma = 0.001$ and $\omega = 0.4$. Our computations were carried out in MATLAB using the package Regularization Tools of Hansen [7].

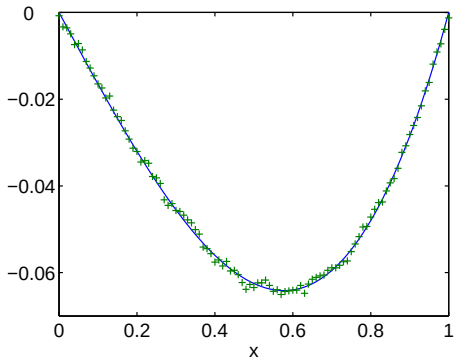


Figure 1: Function $g(x) = (x^3 - x)/6$ and correlated data (+) from the AR(1) model with $n = 101$, $\sigma = 0.001$ and $\omega = 0.4$

As discussed in [16], the generalized eigenvalues $\bar{\lambda}_i$, which satisfy $n^{-1}K^TK\bar{\boldsymbol{\phi}}_i = \bar{\lambda}_i M^T M\bar{\boldsymbol{\phi}}_i$, $i = 1, \ldots, n$, decay like i^{-6} for $i = 1 \ldots, n-2$, and $\bar{\lambda}_{n-1} = \bar{\lambda}_{n-2} = 0$. Since both K and M are $n \times n$ matrices, the results of Section 3 apply with $m = 0$. Because not all the $\bar{\lambda}_i$, $i = 1, \ldots, n$, are positive, Lemmas 3.1 and 3.3 imply that $V'(0) < 0$, $\overline{V}'(0) < 0$ and $\overline{V}_1'(0) < 0$ with probability 1, and so, for this example, GCV, RGCV and R_1GCV will not choose the extreme value 0. However, from Theorem 3.1 and the subsequent discussion, it is quite likely that the GCV estimate will be extremely small if the sample size is small or the errors are red noise with strong correlation.

The simulation results are consistent with the theory. For uncorrelated data and correlated data with $\omega < 0$, GCV gives good and reasonably stable estimates. By contrast, as ω is increased from near 0, the GCV estimates have substantially higher variability, with a greater tendency to have an extremely small value. To illustrate this, Figures 2(a) and (b) show 20 replicates of the GCV function for uncorrelated errors and for correlated errors with $\omega = 0.4$, respectively, where $n = 101$ and $\sigma = 0.001$, together with the corresponding GCV estimates marked with a + symbol. In Figure 2(a) most of the estimates are concentrated between 10^{-5} and 10^{-4}, with only one very small estimate at 10^{-9}. In Figure 2(b), there is considerable variability in the GCV estimates, ranging from 3×10^{-13} to 3×10^{-8}, all of which are too small. In the corresponding plot for $\omega = 0.8$ (not shown), the GCV estimates lie between 10^{-11} and 2×10^{-9}.

For appropriate values of γ, the RGCV and R_1GCV estimates are much more stable than the GCV estimate. We use the same values of γ as in [16, 17], i.e. $\gamma = 0.1$ for RGCV and $\gamma = 0.9999$ for R_1GCV, which give good results for uncorrelated

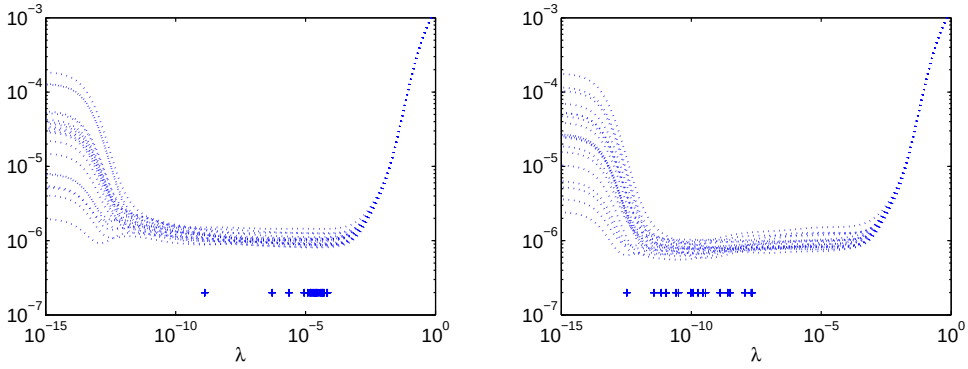


Figure 2: Twenty replicates of the GCV function for (a) uncorrelated data and (b) correlated data with $\omega = 0.4$, and $n = 101$ and $\sigma = 0.001$, together with the corresponding GCV estimates marked with a + symbol

data. Figures 3(a) and (b) show replicates of the RGCV ($\gamma = 0.1$) function and the R_1GCV ($\gamma = 0.9999$) function, respectively, for the same 20 correlated data sets used in Figure 2(b) ($\omega = 0.4$), together with the corresponding RGCV and R_1GCV estimates. Clearly, the RGCV and R_1GCV estimates are much more stable than the GCV estimate. For very strongly correlated data with red noise (e.g. $\omega = 0.8$), the RGCV estimate is also unstable, while the R_1GCV estimate remains stable. This is consistent with Theorem 3.4 and the discussion above it about RGCV.

To compare the GCV, RGCV ($\gamma = 0.1$) and R_1GCV ($\gamma = 0.9999$) estimates, we use the prediction error $R(\lambda) = n^{-1}\|K\boldsymbol{f}_\lambda - K\boldsymbol{f}_0\|^2$ and prediction risk $ER(\lambda)$, as well as the error

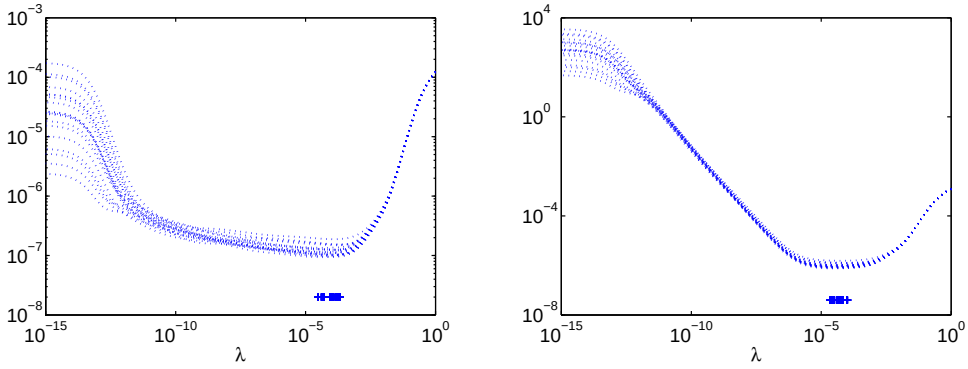
$$R_1(\lambda) = \sum_{i=1}^{n-2} n^{-2}(K\boldsymbol{f}_\lambda - K\boldsymbol{f}_0, \bar{\boldsymbol{\phi}}_i)^2 \bar{\lambda}_i^{-1}, \tag{5.1}$$

defined in [16], and corresponding risk $ER_1(\lambda)$. The error $R_1(\lambda)$ behaves like a squared discrete Sobolev seminorm of order 1 of the error $\boldsymbol{f}_\lambda - \boldsymbol{f}_0$, and, therefore, it is a better measure than $R(\lambda)$ of the accuracy of the regularized solution [16]. Define the inefficiencies I_R and I_{ER} as

$$I_R = R(\lambda)/\min R(\lambda) \quad \text{and} \quad I_{ER} = ER(\lambda)/\min ER(\lambda),$$

and similarly define I_{R_1} and I_{ER_1}. The closer the inefficiency is to 1, the better is the choice λ.

Figure 4 shows box plots of the inefficiencies for the GCV, RGCV and R_1GCV estimates (with GCV (left), RGCV (middle) and R_1GCV (right) in each group of three) corresponding to 200 replicates of the data with $n = 101$, $\sigma = 0.001$ and



(a) (b)

Figure 3: Twenty replicates of the (a) RGCV ($\gamma = 0.1$) function and (b) R_1GCV ($\gamma = 0.9999$) function for correlated data with $\omega = 0.4$, $n = 101$ and $\sigma = 0.001$, together with the corresponding RGCV and R_1GCV estimates marked with a + symbol

$\omega = -0.8, -0.4, 0, 0.4, 0.8$, where 0 denotes the uncorrelated case. The means and medians of the inefficiencies are given in Table 1. Clearly, for uncorrelated data and correlated data with $\omega < 0$ (i.e. white or blue noise), GCV, RGCV and R_1GCV all give good results, though GCV has a significant number of outliers. On the other hand, when $\omega > 0$ (i.e. red noise), RGCV and R_1GCV have much better performance than GCV. In fact, for $\omega = 0.4$ and $\omega = 0.8$, almost all the inefficiencies I_{R_1} and I_{ER_1} for GCV are off the scale (i.e. greater than 50) because of severe undersmoothing. For $\omega = 0.4$, both RGCV and R_1GCV perform very well, and for $\omega = 0.8$, R_1GCV has much better performance than both GCV and RGCV.

Note that the good performance of RGCV and R_1GCV does not require a special choice depending on ω of the robustness parameter γ. The values of γ used for RGCV ($\gamma = 0.1$) and for R_1GCV ($\gamma = 0.9999$), which are close to optimal for uncorrelated data in this example [16, 17], also yield good results for correlated data. Therefore, this one choice of γ for each of RGCV and R_1GCV can be used with reasonable confidence for data with unknown correlation.

The AR(1) model for the correlated errors in this section is different from the covariance assumption used in Section 4, so it appears that the good performance of R_1GCV is not overly sensitive to the form of the covariance. The results presented here for GCV, RGCV and R_1GCV are consistent with those of a large simulation study in [1] involving a range of ill-posed problems with both uncorrelated errors and correlated errors generated by a moving average process.

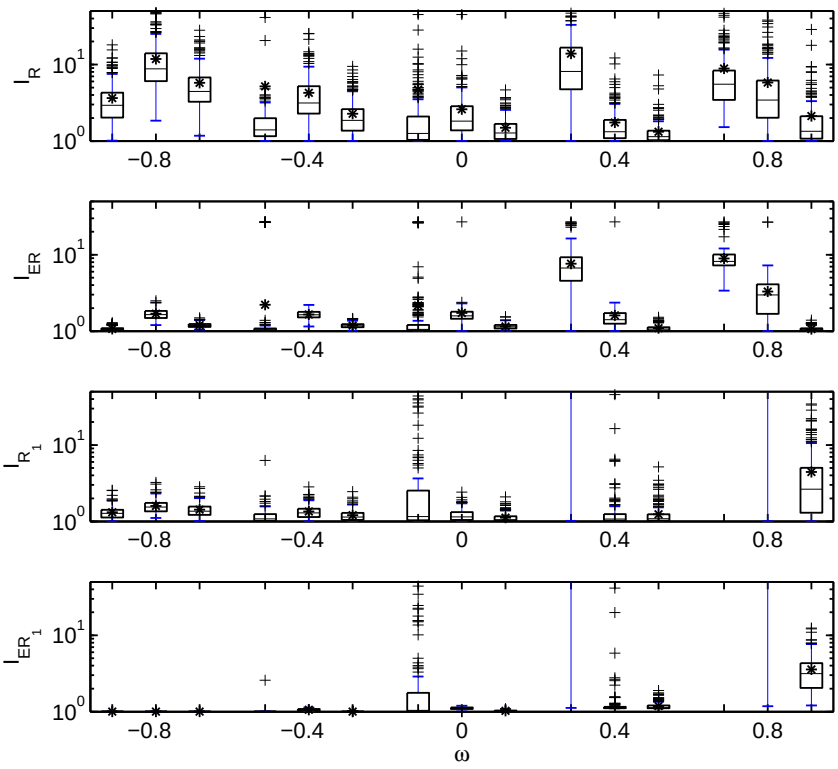


Figure 4: Box plots of inefficiencies I_R, I_{ER}, I_{R_1} and I_{ER_1} in groups of three for GCV (left), RGCV ($\gamma = 0.1$) (middle) and $\mathrm{R_1GCV}$ ($\gamma = 0.9999$) (right), where each group has the same 200 replicates of the data with $n = 101$, $\sigma = 0.001$ and $\omega = -0.8, -0.4, 0$ (uncorrelated), $0.4, 0.8$. In each box plot, the whiskers have maximum length of 4 times the interquartile range and the mean is marked with a $*$ symbol.

References

[1] F. Bauer and M. A. Lukas, Comparing parameter choice methods for regularization of ill-posed problems, submitted (2010).

[2] F. Bauer, P. Mathé and S. Pereverzev, Local solutions to inverse problems in geodesy. The impact of the noise covariance structure upon the accuracy of estimation, J. Geod., **81** (2007), pp. 39-51.

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

	$\omega = -0.8$			$\omega = -0.4$			$\omega = 0$		
	G	RG	R_1G	G	RG	R_1G	G	RG	R_1G
mean(I_R)	3.64	11.84	5.76	5.18	4.28	2.28	4.70	2.61	1.49
med.(I_R)	2.94	8.80	4.44	1.40	3.14	1.87	1.25	1.83	1.27
mean(I_{ER})	1.07	1.69	1.20	2.21	1.66	1.18	2.16	1.73	1.15
med.(I_{ER})	1.05	1.65	1.18	1.04	1.64	1.17	1.05	1.58	1.14
mean(I_{R_1})	1.31	1.59	1.42	10^7	1.35	1.21	10^7	10^6	1.12
med.(I_{R_1})	1.28	1.54	1.36	1.08	1.29	1.14	1.16	1.15	1.05
mean(I_{ER_1})	1.00	1.00	1.00	10^7	1.06	1.00	10^7	10^6	1.02
med.(I_{ER_1})	1.00	1.00	1.00	1.00	1.06	1.00	1.04	1.10	1.02

	$\omega = 0.4$			$\omega = 0.8$		
	G	RG	R_1G	G	RG	R_1G
mean(I_R)	13.93	1.75	1.33	8.89	5.83	2.12
med.(I_R)	8.11	1.32	1.14	5.54	3.43	1.34
mean(I_{ER})	7.64	1.61	1.10	8.92	3.30	1.07
med.(I_{ER})	6.70	1.41	1.06	8.18	2.98	1.05
mean(I_{R_1})	10^6	10^5	1.24	10^6	10^5	4.44
med.(I_{R_1})	10^5	1.08	1.08	10^5	10^3	2.65
mean(I_{ER_1})	10^6	10^5	1.19	10^6	10^5	3.57
med.(I_{ER_1})	10^5	1.13	1.13	10^5	10^3	3.15

Table 1: Means and medians of the inefficiencies I_R, I_{ER}, I_{R_1} and I_{ER_1} for GCV (G), RGCV (RG) and R_1GCV (R_1G) for 200 replicates of the data with $n = 101$, $\sigma = 0.001$ and $\omega = -0.8, -0.4, 0$ (uncorrelated), $0.4, 0.8$

[4] D. J. Cummins, T. G. Filloon and D. Nychka, Confidence intervals for nonparametric curve estimates: Toward more uniform pointwise coverage, J. Amer. Statist. Assoc., **96** (2001), pp. 233-246.

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

[6] C. Han and C. Gu, Optimal smoothing with correlated data, Sankhyā **70-A** (2008), pp. 38-72.

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

[8] P. C. Hansen, Rank-Deficient and Discrete Ill-Posed Problems, SIAM, Philadelphia, 1998.

[9] M. Jansen and A. Bultheel, Multiple wavelet threshold estimation by generalized cross validation for images with correlated noise, IEEE Trans. Image Process., **8** (1999), pp. 947-953.

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

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

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

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

[14] M. A. Lukas, Convergence rates for moment collocation solutions of linear operator equations, Numer. Funct. Anal. Optim., **16** (1995), pp. 743-750.

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

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

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

[18] M. A. Lukas, F. R. de Hoog and R. S. Anderssen, Spline smoothing using robust GCV, Technical Report 08-154, CMIS, CSIRO, 2008.

[19] D. S. Mitrinović, J. E. Pečarić and A. M. Fink, Classical and New Inequalities in Analysis, Kluwer, Dordrecht, 1993.

[20] J. Opsomer, Y. Wang and Y. Yang, Nonparametric regression with correlated errors, Statist. Sci., **16** (2001), pp.134-153.

[21] T. Robinson and R. Moyeed, Making robust the cross-validation choice of smoothing parameter in spline smoothing regression, Comm. Statist. Theory Methods, **18** (1989), pp. 523-539.

[22] G. B. Rybicki and W. H. Press, Class of fast methods for processing irregularly sampled or otherwise inhomogeneous one-dimensional data, Phys. Rev. Lett., **74** (1995), pp. 1060-1063.

[23] A. M. Thompson, J. W. Kay and D. M. Titterington, A cautionary note about crossvalidatory choice, J. Stat. Comput. Simul., **33** (1989), pp. 199-216.

[24] R. Vio, P. Ma, W. Zhong, J. Nagy, L. Tenorio and W. Wamsteker, Estimation of regularization parameters in multiple-image deblurring, Astron. Astrophys., **423** (2004), pp. 1179-1186.

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

[26] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, 1990.

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