Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

CHAPTER 4

Estimating the Smoothing Parameter

4.1 The importance of a good choice of λ.

Figures 4.1, 4.2, and 4.3 from Wahba and Wold (1975) were part of the results of a Monte Carlo study to examine the behavior of ordinary cross validation (OCV) for estimating the smoothing parameter in a cubic smoothing spline. The dashed line in each of these figures is a plot of $f(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x})$, and the dots enclosed in boxes represent values of

$$y_i = f\left(\frac{i}{n}\right) + \epsilon_i, \quad i = 1, \ldots, 100, \tag{4.1.1}$$

where the ϵ_i's come from a random number generator simulating independently and identically distributed $\mathcal{N}(0, \sigma^2)$ random variables, with standard deviation $\sigma = .2$. The solid line in each figure is f_λ, the minimizer of $1/n \sum_{i=1}^{n}(y_i - f(i/n))^2 + \lambda \int_0^1 (f''(x))^2 dx$. In Figure 4.1 λ is too small, in Figure 4.2 too big, and in Figure 4.3 "about right." The parameter λ in Figure 4.3 was estimated by OCV, also known as the "leaving-out-one" method, to be described. Evidently the visual appearance of the picture is quite dependent on λ, which is not surprising as we recall that as λ runs from zero to ∞, f_λ runs from an interpolant to the data, to the straight line best fitting the data in a least squares sense. Figures 4.4–4.8 provide a two-dimensional example. Figure 4.4 gives a plot of a test function used by Franke (1979), this function is a linear combination of four normal density functions. Figure 4.5 gives a schematic plot of the data $y_i = f(x_1(i), x_2(i)) + \epsilon_i$, where the ϵ_i come from a random number generator as before. The (x_1, x_2) take on values on a 7×7 regular grid and the $n = 49$ y_i's have been joined by straight lines in an attempt to make the picture clearer. Figure 4.6 gives a plot of f_λ, the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_1(i), x_2(i)))^2 + \iint_{-\infty}^{\infty} (f_{x_1x_1}^2 + 2f_{x_1x_2}^2 + f_{x_2x_2}^2)\, dx_1\, dx_2$$

with λ evidently too large, Figure 4.7 gives f_λ with λ too small, and Figure 4.8 gives f_λ with λ "about right." Figures 4.4–4.8 are from Wahba (1979c). Generalized cross validation (to be discussed) was used to choose λ in Figure 4.8.

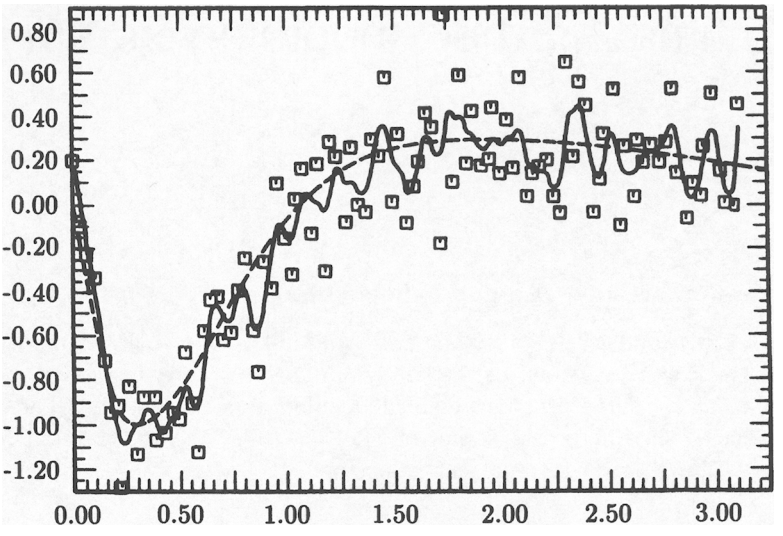

FIG. 4.1. *Data generated according to the model* (4.1.1). *Dashed curve is* $f(x)$. *Solid curve is fitted spline with* λ *too small.*

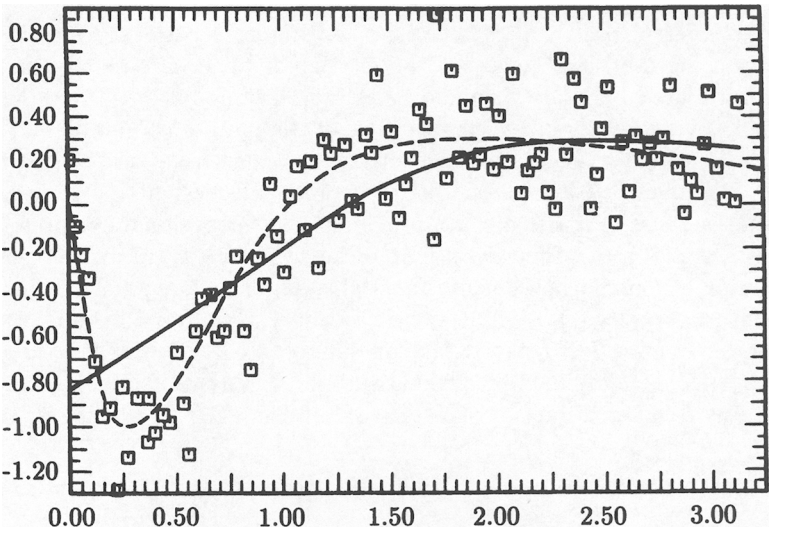

FIG. 4.2. *Same data as in Figure* 4.1. *Spline (solid curve) is fitted with* λ *too big.*

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

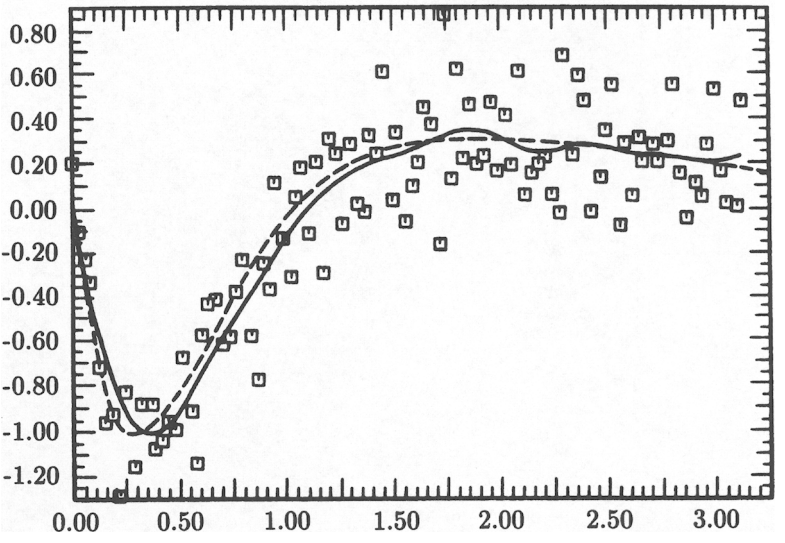

FIG. 4.3. *Same data as in Figure* 4.2. *Spline (solid curve) is fitted with the OCV estimate of* λ.

4.2 Ordinary cross validation and the "leaving-out-one" lemma.

Next we will explain these methods. Ordinary cross validation (OCV) goes as follows. Let $f_\lambda^{[k]}$ be the minimizer of

$$\frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - f(x_i))^2 + \lambda\int_0^1 (f''(u))^2\,du. \tag{4.2.1}$$

Then the "ordinary cross-validation function" $V_0(\lambda)$ is

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - f_\lambda^{[k]}(x_k)\right)^2, \tag{4.2.2}$$

and the OCV estimate of λ is the minimizer of $V_0(\lambda)$. More generally, if we let $f_\lambda^{[k]}$ be the minimizer of

$$\frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - L_i f)^2 + \lambda\|P_1 f\|^2 \tag{4.2.3}$$

(assumed unique), then

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - L_k f_\lambda^{[k]})^2. \tag{4.2.4}$$

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

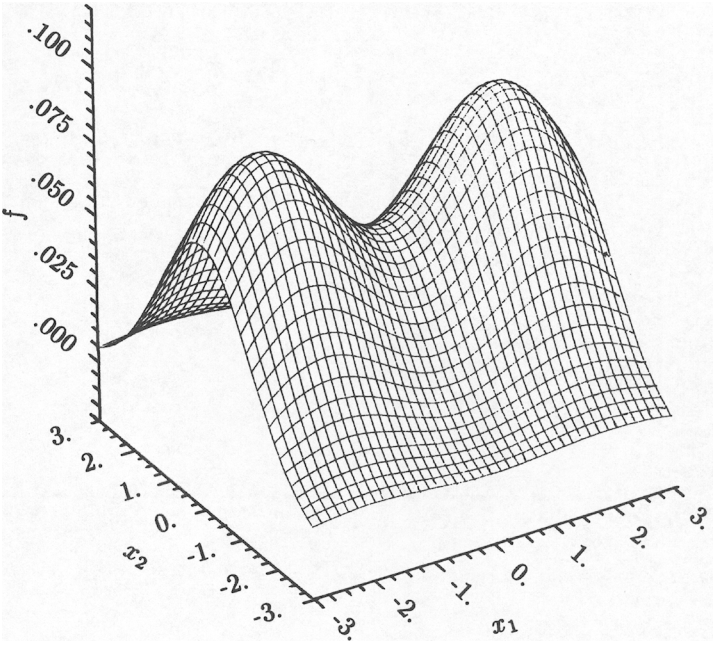

FIG. 4.4. *The actual surface.*

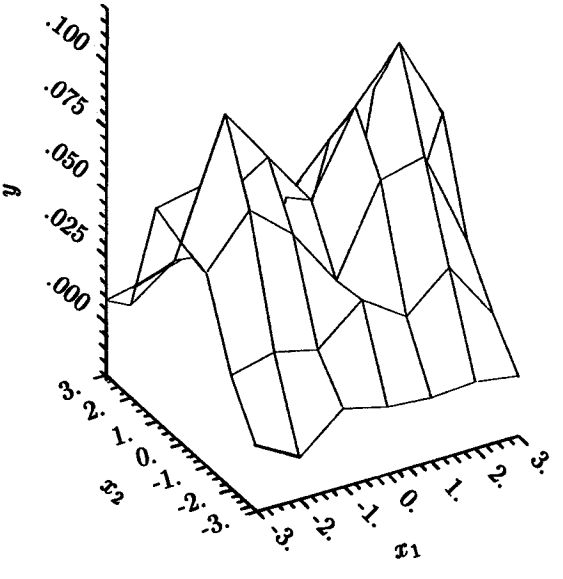

FIG. 4.5. *The data.*

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

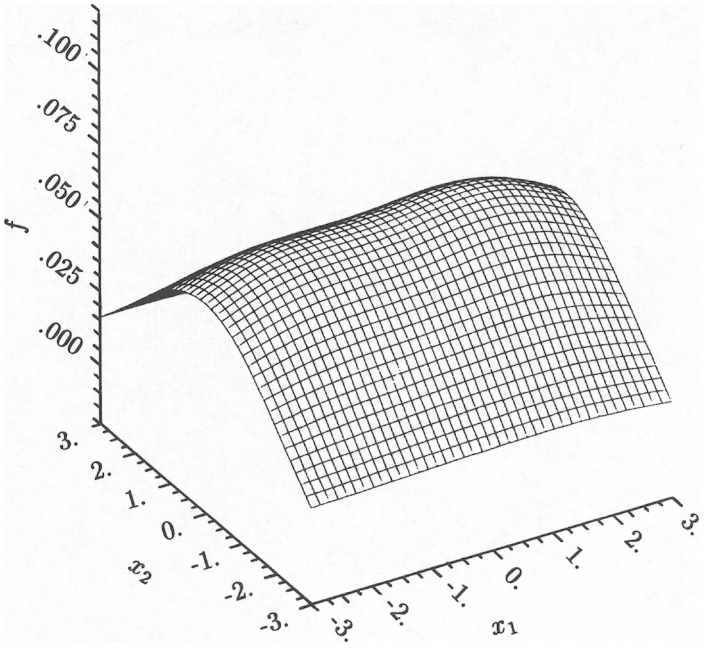

FIG. 4.6. *f_λ with λ too large, $\lambda = 100\hat{\lambda}$.*

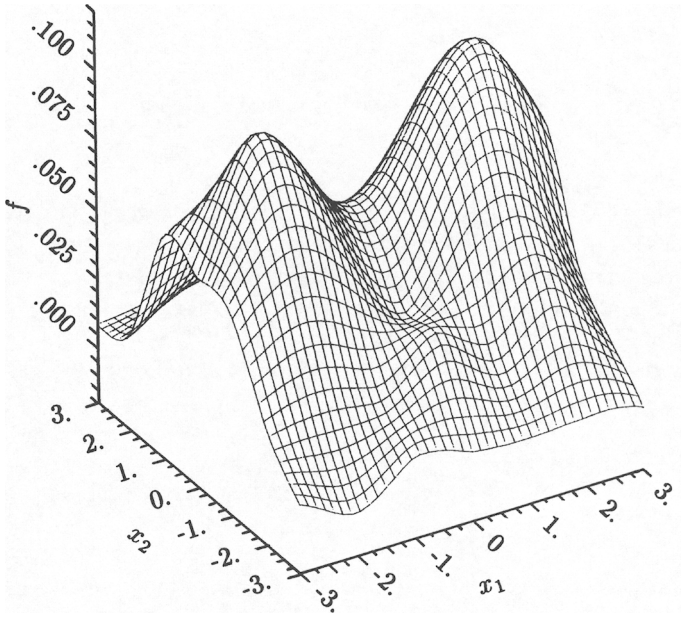

FIG. 4.7. *f_λ with λ too small, $\lambda = .01\hat{\lambda}$.*

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

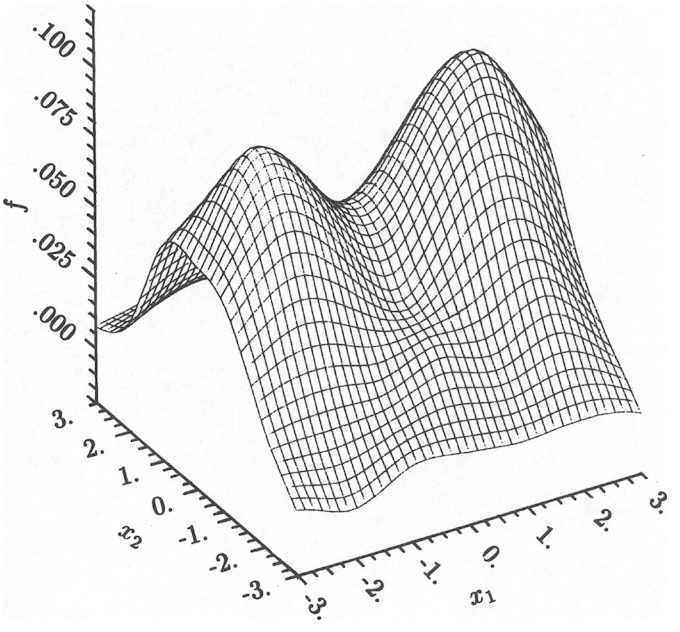

FIG. 4.8. *f_λ with λ estimated by GCV.*

OCV was suggested by Allen (1974) in the context of regression and by Wahba and Wold (1975) in the context of smoothing splines, after hearing Mervyn Stone discuss it in the context of determining the degree of a polynomial in polynomial regression. The idea of leaving out one or several no doubt is quite old (see, e.g., Mosteller and Wallace (1963)).

We now prove the "*leaving-out-one*" lemma (Craven and Wahba (1979)).

LEMMA 4.2.1. *Let $f_\lambda^{[k]}$ be the solution to the following problem. Find $f \in \mathcal{H}_R$ to minimize*

$$\frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - L_i f)^2 + \lambda\|P_1 f\|^2.$$

Fix k and z and let $h_\lambda[k,z]$ be the solution to the following problem. Find $f \in \mathcal{H}_R$ to minimize

$$\frac{1}{n}\left[(z - L_k f)^2 + \sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - L_i f)^2\right] + \lambda\|P_1 f\|^2. \tag{4.2.5}$$

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Then $h_\lambda[k, L_k f_\lambda^{[k]}] = f_\lambda^{[k]}$.

Proof. Let $\tilde{y}_k = L_k f_\lambda^{[k]}$, let $h = f_\lambda^{[k]}$, and let f be any element in $\mathcal{H}_R$ different from h. Then

$$\begin{aligned}
&\frac{1}{n}\left[(\tilde{y}_k - L_k h)^2 + \sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - L_i h)^2\right] + \lambda\|P_1 h\|^2 \\
&= \frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - L_i h)^2 + \lambda\|P_1 h\|^2 \\
&< \frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - L_i f)^2 + \lambda\|P_1 f\|^2 \\
&\leq \frac{1}{n}\left[(\tilde{y}_k - L_k f)^2 + \sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - L_i f)^2\right] + \lambda\|P_1 f\|^2. \qquad (4.2.6)
\end{aligned}$$

Now consider the following identity:

$$y_k - L_k f_\lambda^{[k]} = \frac{(y_k - L_k f_\lambda)}{(1 - a_{kk}^*(\lambda))} \qquad (4.2.7)$$

where

$$a_{kk}^* = \frac{L_k f_\lambda - L_k f_\lambda^{[k]}}{y_k - L_k f_\lambda^{[k]}}. \qquad (4.2.8)$$

By the leaving-out-one lemma, and by letting $\tilde{y}_k = L_k f_\lambda^{[k]}$ and noting that $L_k f_\lambda = L_k h_\lambda[k, y_k]$ by definition, we can write

$$a_{kk}^* = \frac{L_k h_\lambda[k, y_k] - L_k h_\lambda[k, \tilde{y}_k]}{y_k - \tilde{y}_k}. \qquad (4.2.9)$$

Thus, looking at $L_k f_\lambda$ as a function of the kth data point, we see that $a_{kk}^*(\lambda)$ is nothing more than a divided difference of this function taken at y_k and $\tilde{y}_k$. However, $L_k f_\lambda$ is linear in each data point, so we can replace this divided difference by a derivative. Thus, we have shown that

$$a_{kk}^*(\lambda) = \frac{\partial L_k f_\lambda}{\partial y_k} = a_{kk}(\lambda), \qquad (4.2.10)$$

where $a_{kk}(\lambda)$ is the kkth entry of the influence matrix $A(\lambda)$, given in (1.3.23).

Thus, we have the following OCV *identity*.

THEOREM 4.2.1.

$$\frac{1}{n}\sum_{k=1}^{n}(y_k - L_k f_\lambda^{[k]})^2 \equiv V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - L_k f_\lambda)^2/(1 - a_{kk}(\lambda))^2. \qquad (4.2.11)$$

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Later, we will generalize the optimization problem of (1.3.4) as follows. Find $f \in \mathcal{C} \subset \mathcal{H}_R$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - N_i f)^2 + \lambda\|P_1 f\|^2 \tag{4.2.12}$$

where $\mathcal{C}$ is some closed convex set in $\mathcal{H}_R$, and $N_i f$ is a (possibly) nonlinear functional. Suppose that the N_i and $\mathcal{C}$ are such that (4.2.12) has a unique minimizer in $\mathcal{C}$, as does (4.2.12) with the kth term deleted. Then it is easy to see that the inequalities of (4.2.6) still hold, with L_i replaced by N_i. Therefore, if $f_\lambda^{[k]}$ is the minimizer in $\mathcal{C}$ of (4.2.12) with the kth term deleted, and $h_\lambda[k,z]$ is the minimizer in $\mathcal{C}$ of (4.2.12) with y_k replaced by z, then, as before, $h_\lambda[k, N_k f_\lambda^{[k]}] = f_\lambda^{[k]}$. Thus the OCV identity generalizes to

$$\frac{1}{n}\sum_{k=1}^{n}(y_k - N_k f_\lambda^{[k]})^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k - N_k f_\lambda)^2/(1 - a_{kk}^*(\lambda))^2 \tag{4.2.13}$$

where

$$a_{kk}^* = \frac{N_k h_\lambda[k, y_k] - N_k h_\lambda[k, \tilde{y}_k]}{y_k - \tilde{y}_k} \tag{4.2.14}$$

with $\tilde{y}_k = N_k f_\lambda^{[k]}$. Now, however, $(\partial N_k f_\lambda / \partial y_k)|_{y_k}$, if it exists, will only be, in general, an approximation to $a_{kk}^*(\lambda)$ of (4.2.14).

4.3 Generalized cross validation.

Generalized cross validation (GCV) for the problem of (1.3.4) is obtained by replacing $a_{kk}(\lambda)$ by $\mu_1(\lambda) = 1/n\sum_{i=1}^{n} a_{ii}(\lambda) = 1/n \operatorname{Tr} A(\lambda)$. The GCV function $V(\lambda)$ is defined by

$$\begin{aligned} V(\lambda) &= \frac{1}{n}\sum_{k=1}^{n}(y_k - L_k f_\lambda)^2/(1 - \mu_1(\lambda))^2 \\ &\equiv \frac{1}{n}\|(I - A(\lambda))y\|^2 / \left[\frac{1}{n}\operatorname{Tr}(I - A(\lambda))\right]^2. \end{aligned} \tag{4.3.1}$$

$V(\lambda)$ may be viewed as a weighted version of $V_0(\lambda)$, since

$$V(\lambda) = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - L_k f_\lambda^{[k]}\right)^2 w_{kk}(\lambda)$$

where $w_{kk}(\lambda) = (1 - a_{kk}(\lambda))^2/(1 - \mu_1(\lambda))^2$. If $a_{kk}(\lambda)$ is independent of k, then $V_0(\lambda) \equiv V(\lambda)$. The "generalized" version was an attempt to achieve certain desirable invariance properties that do not generally hold for ordinary cross validation. Let Γ be any $n \times n$ orthogonal matrix, and consider a new data vector $\tilde{y} = \Gamma y$ and a new set of bounded linear functionals $(\tilde{L}_1, \ldots, \tilde{L}_n)' = \Gamma(L_1, \ldots, L_n)'$. The problem of estimating f from data

$$\tilde{y}_i = \tilde{L}_i f + \tilde{\epsilon}_i \ , \ i = 1, \ldots, n,$$

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where $\tilde{\epsilon} = \Gamma\epsilon$ is the same as the problem of estimating f from

$$y_i = L_i f + \epsilon_i, \ i = 1, \ldots, n,$$

since $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$. However, it is not hard to see that, in general, OCV can give a different value of λ. The GCV estimate of λ is invariant under this transformation.

The original argument by which GCV was obtained from OCV can be described most easily with regards to a ridge regression problem (see Golub, Heath, and Wahba (1979)). Let

$$y = X\beta + \epsilon, \tag{4.3.2}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, and to avoid irrelevant discussion suppose X is $n \times n$. β will be estimated as the minimizer of

$$\frac{1}{n}\|y - X\beta\|^2 + \lambda\, \beta'\beta.$$

Let the singular value decomposition (see Dongarra et al. (1979)) of X be UDV', and write

$$\tilde{y} = D\gamma + \tilde{\epsilon} \tag{4.3.3}$$

where $\tilde{y} = U'y$, $\gamma = V'\beta$, and $\tilde{\epsilon} = U'\epsilon$. The problem is invariant under this transformation. On the other hand, since

$$\tilde{y}_i = d_i\gamma_i + \tilde{\epsilon}_i \ , \ i = 1, \ldots, n, \tag{4.3.4}$$

where d_i is the iith entry of D, it is fairly clear that a leaving-out-one method of choosing λ is not going to work too well since the rows are uncoupled. (In fact, $V_0(\lambda)$ is independent of λ!) However, if the singular values of D come in pairs, there is an orthogonal matrix W for which WDW' is a symmetric circulant matrix, and symmetric circulant matrices may be viewed as having rows that are maximally coupled. Recall that a circulant matrix has the property that if the first row is $(\theta_0, \theta_1, \ldots, \theta_{n-1})$, then the jth row is $(\theta_{n-j}, \ldots, \ \theta_{n-1}, \theta_0, \theta_1, \ldots, \theta_{n-j-1})$. The matrix W is the discrete Fourier transform matrix and W and D are given, for even n, in Table 4.1. Transforming (4.3.3) by W gives

$$z = WDW'\delta + \xi \tag{4.3.5}$$

where $z = W\tilde{y}$, $\delta = W\gamma$, $\xi = W\epsilon$, and WDW' is circulant. Intuitively, the design matrix D has "maximally uncoupled" rows, while the design matrix WDW', being circulant, has "maximally coupled" rows. The influence matrix $A(\lambda)$ for the problem (4.3.5) is circulant and hence constant down the diagonal. GCV is equivalent here to transforming the original problem (4.3.2) into the "maximally coupled" form (4.3.5), doing OCV, and transforming back.

The GCV estimate $\hat{\lambda}$ of λ is known to have a number of favorable properties, both from practical experience and theoretically. For some Monte Carlo experimental results, see, e.g., Craven and Wahba (1979), Merz (1980), Nychka

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 4.1
The discrete Fourier transform matrix W and corresponding diagonal matrix D.

$$W = \begin{pmatrix} -c_0 \\ -\sqrt{2}c_1 \\ \vdots \\ -\sqrt{2}c_{n/2-1} \\ -c_{n/2} \\ -\sqrt{2}s_1 \\ \vdots \\ -\sqrt{2}s_{n/2-1} \end{pmatrix}$$

where

$$\begin{aligned} c_0 &= \frac{1}{\sqrt{n}}(1,\ldots,1), \\ c_\nu &= \frac{1}{\sqrt{n}}(\cos 2\pi\nu\frac{1}{n}, \cos 2\pi\nu\frac{2}{n}, \ldots, \cos 2\pi\nu\frac{n}{n}), \\ s_\nu &= \frac{1}{\sqrt{n}}(\sin 2\pi\nu\frac{1}{n}, \sin 2\pi\nu\frac{2}{n}, \ldots, \sin 2\pi\nu\frac{n}{n}). \end{aligned}$$

$$D = \begin{pmatrix} d_0 & & & & & & & 0 \\ & d_1 & & & & & & \\ & & \ddots & & & & & \\ & & & d_{n/2-1} & & & & \\ & & & & \frac{1}{2}d_{n/2} & & & \\ & & & & & d_1 & & \\ & & & & & & \ddots & \\ 0 & & & & & & & d_{n/2-1} \end{pmatrix}$$

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

et al. (1984), Vogel (1986), Shahrary and Anderson (1989), Scott and Terrell (1987), Woltring (1985), the rejoinder in Hardle, Hall, and Marron (1988), etc. The so-called "weak cross-validation theorem" was proposed and nearly proved for the smoothing spline case in Craven and Wahba (1979). Utreras (1978, 1981b) completed the proof by obtaining rigorously certain properties of some eigenvalues necessary to complete the proof. Properties of eigenvalues in other cases were obtained by Utreras (1979, 1981b), Cox (1983), and others. See also Wahba (1977a). Strong theorems were obtained by Speckman (1985), Li(1985, 1986, 1987). Generalizations are discussed in Hurvich (1985), O'Sullivan (1986a), Altman (1987), Gu (1990), Friedman and Silverman (1989). The arguments below are adapted from Craven and Wahba (1979) and Wahba (1985e).

4.4 Properties of the GCV Estimate of λ.

GCV is a *predictive mean-square error* criteria, which is not surprising given its source. Define the predictive mean-square error $T(\lambda)$ as

$$T(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(L_i f_\lambda - L_i f)^2. \tag{4.4.1}$$

The GCV estimate of λ is an estimate of the minimizer of $T(\lambda)$. $T(\lambda)$ depends on the unknown f as well as the unknown $\epsilon_1, \ldots, \epsilon_n$. The expected value of $T(\lambda)$, $ET(\lambda)$ is given by

$$ET(\lambda) = E\frac{1}{n}\sum_{k=1}^{n}(L_k f_\lambda - L_k f)^2.$$

Letting $g = (L_1 f, \ldots, L_n f)'$ we have $(L_1 f_\lambda, \cdots, L_n f_\lambda)' = A(\lambda)y = A(\lambda)(g+\epsilon)$, and

$$\begin{aligned} ET(\lambda) &= \frac{1}{n}E\|A(\lambda)(g+\epsilon) - g\|^2 \\ &= \frac{1}{n}\|(I - A(\lambda))g\|^2 + \frac{\sigma^2}{n}\operatorname{Tr} A^2(\lambda) \\ &= b^2(\lambda) + \sigma^2\mu_2(\lambda), \quad \text{say.} \end{aligned}$$

These terms are known as the bias and variance terms, respectively. Using the representation for $I - A(\lambda)$ given in (1.3.23),

$$I - A(\lambda) = n\lambda Q_2(Q_2'(\Sigma + n\lambda I)Q_2)^{-1}Q_2',$$

letting the eigenvector eigenvalue decomposition of $Q_2'\Sigma Q_2$ be UDU', where U is $(n-M)\times(n-M)$ orthogonal and D is diagonal and $\Gamma = Q_2\, U$, we have

$$I - A(\lambda) = n\lambda\Gamma(D + n\lambda I)^{-1}\Gamma'. \tag{4.4.2}$$

Letting

$$h = \Gamma' g$$

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

we have

$$b^2(\lambda) = \frac{1}{n}\sum_{\nu=1}^{n-M}\left(\frac{n\lambda h_{\nu n}}{\lambda_{\nu n}+n\lambda}\right)^2,$$
$$\mu_2(\lambda) = \frac{1}{n}\left(\sum_{\nu=1}^{n-M}\left(\frac{\lambda_{\nu n}}{\lambda_{\nu n}+n\lambda}\right)^2 + M\right) \tag{4.4.3}$$

where $h_{\nu n}$, $\nu = 1,\ldots,n-M$ are the components of h, and $\lambda_{\nu n}$ are the diagonal entries of D. If f is in the null space of P_1, that is, f is of the form

$$f = \sum_{\nu=1}^{M}\theta_\nu\phi_\nu,$$

then $g = (L_1 f,\ldots,L_n f)' = T\theta$, where $\theta = (\theta_1,\ldots,\theta_M)'$, and then $h = \Gamma' g = U'Q_2' g = U'Q_2'T\theta = 0$, by the construction of Q_2 in (1.3.18). Then $\mu_2(\lambda)$ is a monotone decreasing function of λ and is minimized for $\lambda = \infty$, which corresponds to f_∞ being the least squares regression of the data on span $\{\phi_1,\ldots,\phi_M\}$, with $\mu_2(\infty) = M/n$. If $\sum_{\nu=1}^{n-M} h_{\nu n}^2 > 0$, then $b^2(\lambda)$ is a monotone increasing function of λ, with $(d/d\lambda)b^2(\lambda)|_{\lambda=0} = 0$, while $\mu_2(\lambda)$ is a monotone decreasing function of λ with strictly negative derivative at $\lambda = 0$, so that $ET(\lambda)$ will have (at least) one minimizer $\lambda^* > 0$.

The "weak GCV theorem" says that there exists a sequence (as $n \to 0$) of minimizers $\tilde{\lambda}$ of $EV(\lambda)$ that comes close to achieving the minimum value of $\min_\lambda ET(\lambda)$. That is, let the expectation inefficiency I^* be defined by

$$I^* = \frac{ET(\tilde{\lambda})}{ET(\lambda^*)}.$$

Then, under some general circumstances to be discussed, $I^* \downarrow 1$ as $n \to \infty$.

We will outline the argument. First,

$$EV(\lambda) = \frac{b^2(\lambda) + \sigma^2(1 - 2\mu_1(\lambda) + \mu_2(\lambda))}{(1-\mu_1(\lambda))^2}, \tag{4.4.4}$$

where

$$\mu_1(\lambda) = \frac{1}{n}\left[\sum_{\nu=1}^{n-M}\frac{\lambda_{\nu n}}{\lambda_{\nu n}+n\lambda} + M\right]. \tag{4.4.5}$$

As before, if $\|P_1 f\|^2 = 0$, then $b^2(\lambda) = 0$, and

$$EV(\lambda) = \frac{\sigma^2}{n}\sum_{\nu=1}^{n-M}\left(\frac{n\lambda}{\lambda_{\nu n}+n\lambda}\right)^2 \Big/ \left(\frac{1}{n}\sum_{\nu=1}^{n-M}\frac{n\lambda}{\lambda_{\nu n}+n\lambda}\right)^2, \tag{4.4.6}$$

which is minimized for $\lambda = \infty$, the same as for $ET(\lambda)$, so in this case $I^* = 1$.

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

We now proceed to the general case. First, some algebraic manipulations give

$$\frac{ET(\lambda) - (EV(\lambda) - \sigma^2)}{ET(\lambda)} = \frac{-\mu_1(2-\mu_1)}{(1-\mu_1)^2} + \frac{\sigma^2}{b^2 + \sigma^2\mu_2} \cdot \frac{\mu_1^2}{(1-\mu_1)^2}$$

and so

$$\frac{|ET(\lambda) - (EV(\lambda) - \sigma^2)|}{ET(\lambda)} \leq h(\lambda) \tag{4.4.7}$$

where

$$h(\lambda) = \left[2\mu_1(\lambda) + \frac{\mu_1^2(\lambda)}{\mu_2(\lambda)}\right] \frac{1}{(1-\mu_1(\lambda))^2}. \tag{4.4.8}$$

Now, using the fact that $\mu_2(\lambda) \geq \mu_1^2(\lambda)$, for all λ, it follows that $EV(\lambda) \geq \sigma^2$, so that (4.4.7) gives

$$ET(\lambda)(1 - h(\lambda)) \leq EV(\lambda) - \sigma^2 \leq ET(\lambda)(1 + h(\lambda)), \text{ for all } \lambda. \tag{4.4.9}$$

Letting λ^* be the minimizer of $ET(\lambda)$, we obtain

$$EV(\lambda^*) - \sigma^2 \leq ET(\lambda^*)(1 + h(\lambda^*))$$

and there must be (at least) one minimizer $\tilde{\lambda}$ of $EV(\lambda)$ in the nonempty set $\Lambda = \{\lambda : EV(\lambda) - \sigma^2 \leq EV(\lambda^*) - \sigma^2\}$ (see Figure 4.9). Thus

$$ET(\tilde{\lambda})(1 - h(\tilde{\lambda})) \leq EV(\tilde{\lambda}) - \sigma^2 \leq EV(\lambda^*) - \sigma^2 \leq ET(\lambda^*)(1 + h(\lambda^*)). \tag{4.4.10}$$

Then (provided $h(\tilde{\lambda}) < 1$),

$$\frac{ET(\tilde{\lambda})}{ET(\lambda^*)} \leq \frac{1 + h(\lambda^*)}{1 - h(\tilde{\lambda})}, \tag{4.4.11}$$

and, if $h(\lambda^*)$ and $h(\tilde{\lambda}) \to 0$, then

$$\frac{ET(\tilde{\lambda})}{ET(\lambda^*)} \downarrow 1. \tag{4.4.12}$$

$h(\lambda^*)$ and $h(\tilde{\lambda})$ will tend to zero if $\mu_1(\lambda^*), \mu_1(\tilde{\lambda}), \mu_1^2(\lambda^*)/(\mu_2(\lambda^*))$ and $\mu_1^2(\tilde{\lambda})/(\mu_2(\tilde{\lambda}))$ tend to zero.

In many interesting cases the eigenvalues $\lambda_{\nu n}$ behave roughly as do $n\nu^{-q}$ for some real number $q > 1$, and the expressions

$$\begin{aligned} \mu_\tau(\lambda) &= \frac{1}{n}\sum_{\nu=1}^{n-M}\left(\frac{\lambda_{\nu n}}{\lambda_{\nu n} + n\lambda}\right)^\tau \simeq \frac{1}{n}\sum \frac{1}{(1+\lambda\nu^q)^\tau}, \quad \tau = 1, 2 \\ &\simeq \frac{1}{n}\int \frac{1}{(1+\lambda x^q)^\tau} \simeq \frac{c_{\tau q}}{n\lambda^{1/q}}, \quad \tau = 1, 2, \quad q > 1 \end{aligned}$$

are valid to the accuracy needed in the proofs.

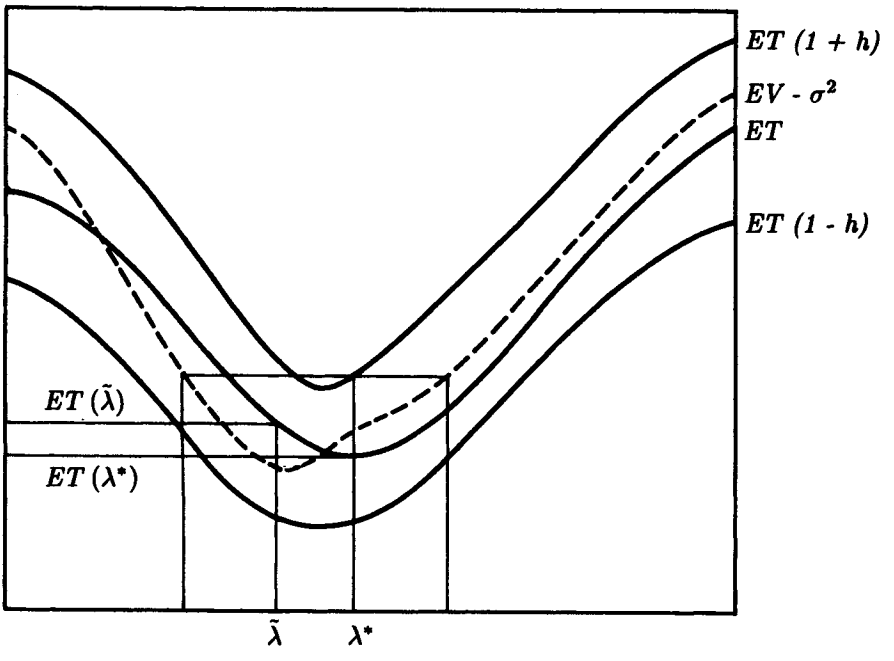

FIG. 4.9. *Graphical suggestion of the proof of the weak* GCV *theorem.*

We only give the reader a very crude argument in support of (4.4.12) and refer the reader to Cox (1988), Utreras (1978) for more rigorous results. The crude argument, for roughly equally spaced evaluation functionals, goes as follows. Suppose

$$R(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu \Phi_\nu(s)\Phi_\nu(t),$$

consider the matrix $\sum$ with ijth entry

$$\begin{aligned} R(t_i,t_j) &= \sum_{\nu=1}^{\infty} \lambda_\nu \Phi_\nu(t_i)\Phi_\nu(t_j) \\ &\approx \sum_{\nu=1}^{n} n\lambda_\nu \frac{\Phi_\nu(t_i)}{\sqrt{n}} \frac{\Phi_\nu(t_j)}{\sqrt{n}}. \end{aligned}$$

If

$$\frac{1}{n}\sum_{l=1}^{n} \Phi_\nu(t_l)\Phi_\mu(t_l) \simeq \int \Phi_\nu(s)\Phi_\mu(s)ds = \delta_{\mu,\nu} \qquad (4.4.13)$$

then roughly $(1/\sqrt{n}\Phi_\nu(t_1),\ldots,1/\sqrt{n}\Phi_\nu(t_n))'$, $\nu = 1,\ldots,n$ are the eigenvectors of Σ and (again roughly), $n\lambda_\nu$, $\nu = 1,2,\ldots$ are the eigenvalues of Σ. The asymptotic behavior of the eigenvalues of $Q_2'\Sigma Q_2$ does not differ "much" from the asymptotic behavior of the eigenvalues of Σ. In particular, if $\alpha_1 \geq \ldots \geq \alpha_n$ are the eigenvalues of Σ and $\lambda_{1n},\ldots,\lambda_{n-M,n}$ are the eigenvalues of $Q_2'\Sigma Q_2$, then, by the variational definition of eigenvalues,

$$\alpha_1 \geq \lambda_{1,n} \geq \alpha_{M+1}$$

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

$$\alpha_2 \geq \lambda_{2,n} \geq \alpha_{M+2}$$

$$\vdots$$

$$\alpha_{n-M} \geq \lambda_{n-M,n} \geq \alpha_n.$$

For the reproducing kernel of W_m^0 (per) of (2.1.4) it is easy to see that (4.4.13) holds (exactly, for $t_l = l/n$), and $\lambda_{\nu n} \simeq n(2\pi\nu)^{-2m}$. If R is a Green's function for a linear differential operator, then the eigenvalues of R can be expected to behave as do the inverses of the eigenvalues of the linear differential operator (see Naimark (1967)).

To study $b^2(\lambda)$, we have the lemma:

$$b^2(\lambda) \leq \lambda \|P_1 f\|^2. \tag{4.4.14}$$

The proof follows upon letting $g = (L_1 f, \ldots, L_n f)'$ and noting that $A(\lambda)g = (L_1 f_\lambda^*, \ldots, L_n f_\lambda^*)'$, where f_λ^* is the solution to the following problem. Find $h \in \mathcal{H}_R$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(g_i - L_i h)^2 + \lambda \|P_1 h\|^2.$$

Then

$$\begin{aligned}
&\frac{1}{n}\|(I - A(\lambda))g\|^2 + \lambda \|P_1 f_\lambda^*\|^2 \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}(g_i - L_i f_\lambda^*)^2 + \lambda \|P_1 f_\lambda^*\|^2 \\
&\quad \leq \frac{1}{n}\sum_{i=1}^{n}(g_i - L_i f)^2 + \lambda \|P_1 f\|^2 \\
&\quad = \lambda \|P_1 f\|^2.
\end{aligned}$$

If $\mu_2(\lambda) \simeq O(1/n\lambda^{1/q})$, then

$$ET(\lambda) \leq O(\lambda) + O\left(\frac{1}{n\lambda^{1/q}}\right) \tag{4.4.15}$$

and thus $ET(\lambda) \downarrow 0$ provided $\lambda \to 0$ and $n\lambda^{1/q} \to \infty$. Furthermore, it can be argued that if λ does not tend to zero (and of course if $n\lambda^{1/q}$ does not tend to infinity), then $ET(\lambda)$ cannot tend to zero. Thus $\mu_T(\lambda^*) \to 0$. Now $EV(\tilde{\lambda}) \downarrow \sigma^2$, since $EV(\tilde{\lambda}) - \sigma^2 \leq ET(\lambda^*)(1 + h(\lambda^*)) \to 0$. If $\sum h_{\nu n}^2 > 0$ it is necessary that $\tilde{\lambda} \to 0$, $n\tilde{\lambda}^{1/q} \to \infty$ in order that $EV(\tilde{\lambda}) \downarrow \sigma^2$ so that the following can be concluded:

$$I^* \downarrow 1.$$

Figure 4.10 gives a plot of $T(\lambda)$ and $V(\lambda)$ for the test function and experimental data of Figures 4.4 and 4.5. It can be seen that $V(\lambda)$ roughly behaves as does $T(\lambda) + \sigma^2$ in the neighborhood of the minimizer of $T(\lambda)$. The

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

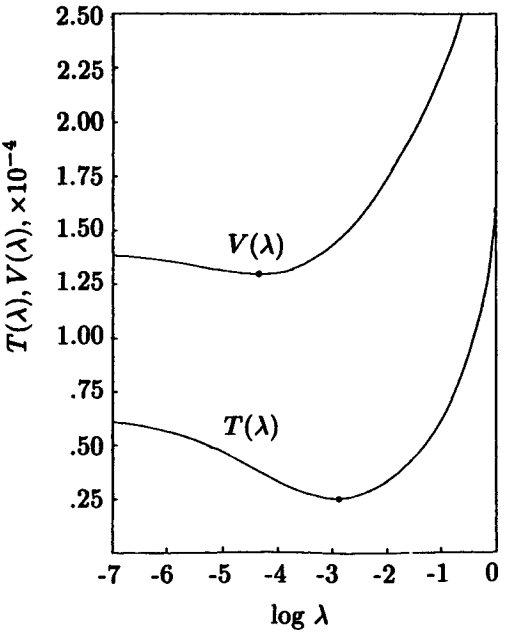

FIG. 4.10. *The mean square error $T(\lambda)$ and the cross validation function $V(\lambda)$.*

sample size was $n = 49$ here. This behavior of V relative to T generally becomes more striking in Monte Carlo experiments as n gets large.

We remark that the parameter m in the penalty functional for polynomial and thin-plate splines can also be estimated by GCV (see Wahba and Wendelberger (1980), Gamber (1979)). V is minimized for each fixed value of m and then the m with the smallest $V(\hat{\lambda})$ is selected.

4.5 Convergence rates with the optimal λ.

It can be seen from (4.4.15) that if $\mu_2(\lambda) = O(1/n\lambda^{1/q})$ and λ is taken as $O(1/n^{q/(q+1)})$, then $ET(\lambda^*) \leq O(1/n^{q/(q+1)})$. If additional conditions hold on the $L_i f$ (that is, on the sequence $(L_{1n}f, \ldots, L_{nn}f)$, $n = 1, 2, \ldots$) then higher rates of convergence can be obtained.

Considering (4.4.3) for the bias, we have

$$\begin{aligned} b^2(\lambda) &= \frac{1}{n}\sum_{\nu=1}^{n-M}\left(\frac{n\lambda h_{\nu n}}{\lambda_{\nu n}+n\lambda}\right)^2 \\ &= \lambda^p \sum_{\nu=1}^{n-M}\left(\frac{n\lambda}{\lambda_{\nu n}+n\lambda}\right)^{2-p}\frac{(h_{\nu n}^2/n)}{(\lambda_{\nu n}/n+\lambda)^p} \\ &\leq \lambda^p \sum_{\nu=1}^{n-M}\frac{(h_{\nu n}^2/n)}{(\lambda_{\nu n}/n)^p} \text{ for any } p \in [0,2]. \end{aligned} \tag{4.5.1}$$

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

If $\sum_{\nu=1}^{n-M}((h_{\nu n}^2/n)/(\lambda_{\nu n}/n)^p)$ is bounded as $n \to \infty$ for some p in $(1, 2]$, then

$$b^2(\lambda) \leq O(\lambda^p) \tag{4.5.2}$$

as $\lambda \to 0$, and if $\mu_2(\lambda) = O(1/n\lambda^{1/q})$, then

$$ET(\lambda) \leq O(\lambda^p) + O\left(\frac{1}{n\lambda^{1/q}}\right),$$

and upon taking $\lambda = O(1/n^{q/(pq+1)})$ we have

$$ET(\lambda^*) \leq O\left(\frac{1}{n^{pq/(pq+1)}}\right). \tag{4.5.3}$$

We know from (4.4.14) that (4.5.2) always holds for $p = 1$; we show this fact another way, by proving that $\sum_{\nu=1}^{n-M}(h_{\nu n}^2/\lambda_{\nu n}) \leq \|P_1 f\|^2$. Letting f_0 be that element in $\mathcal{H}_R$ that minimizes $\|P_1 f_0\|^2$ subject to

$$L_i f_0 = L_i f \equiv g_i, \text{ say},$$

we have, using the calculations in Chapter 1, with $y_i = g_i$ and $\lambda = 0$, that

$$f_0 = \sum_{i=1}^{n} c_i^0 \xi_i + \sum_{\nu=1}^{M} d_\nu^0 \phi_\nu$$

where c^0 and d^0 satisfy $\Sigma c^0 + T d^0 = g$, $T' c^0 = 0$, and so $c^0 = Q_2(Q_2'\Sigma Q_2)^{-1}Q_2' g$. Now $\|P_1 f\|^2 \geq \|P_1 f_0\|^2 = c^{0\prime}\Sigma c^0 = g' Q_2(Q_2'\Sigma Q_2)^{-1}Q_2' g = \sum_{\nu=1}^{n-M} h_{\nu n}^2/\lambda_{\nu n}$.

An example to show when (4.5.1) holds for some $p > 1$ goes as follows. In the case $\mathcal{H}_R = W_m$ (per) with $L_i f = f(i/n)$, $f(t) = \sum f_\nu \Phi_\nu(t)$ (the Φ_ν are sines and cosines here), then $h_{\nu n} \approx \sqrt{n} f_\nu$ by the same argument as that leading up to (4.4.13), $\lambda_{\nu n} \simeq n(2\pi\nu)^{-2m}$, and if $\int_0^1 (f^{(pm)}(x))^2\,dx < \infty$, then

$$\infty > \sum_{\nu=1}^{\infty} (2\pi\nu)^{2pm} f_\nu^2 = \int_0^1 (f^{(pm)}(x))^2\,dx \simeq \sum_{\nu=1}^{n-m} \frac{h_{\nu n}^2/n}{(\lambda_{\nu n}/n)^p}(1 + o(1)).$$

Let

$$R^p(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu^p \Phi_\nu(s)\Phi_\nu(t)$$

for some $p \in (1, 2]$. $f \in \mathcal{H}_{R^p}$ if and only if

$$\sum_{\nu=1}^{\infty} \frac{f_\nu^2}{\lambda_\nu^p} < \infty \tag{4.5.4}$$

where

$$f_\nu = \int f(t)\Phi_\nu(t)\,dt.$$

A general argument similar to that surrounding (4.4.13), (4.5.2), and (4.5.3) would suggest that if the L_i's are roughly uniformly spaced evaluation functionals and $f \in \mathcal{H}_{R^p}$, and $\lambda_\nu = O(\nu^{-q})$, then convergence rates of $O(1/n^{pq/(pq+1)})$ are available. For more general L_i, see Wahba (1985e). Convergence rates for smoothing splines under various assumptions have been found by many authors. See, e.g., Davies and Anderssen (1985), Cox (1983, 1984, 1988), Craven and Wahba (1979), Johnstone and Silverman (1990), Lukas (1981), Ragozin (1983), Rice and Rosenblatt (1983), Silverman (1982), Speckman (1985), Utreras (1981b), Wahba (1977a), and Wahba and Wang (1987).

4.6 Other estimates of λ similar to GCV.

We remark that a variety of criteria $C(\lambda)$ have been proposed such that $\tilde{\lambda}$ is estimated as the minimizer of $C(\lambda)$, where $C(\lambda)$ is of the form

$$C(\lambda) = \|(I - A(\lambda))y\|^2 c(\lambda) \tag{4.6.1}$$

where $c(\lambda) = 1 + 2\mu_1(\lambda) + o(\mu_1(\lambda))$ when $\mu_1 \to 0$ (see Hardle, Hall, and Marron (1988)). Such estimates will have a sequence of minimizers that satisfy the weak GCV theorem.

Note that

$$V(\lambda) = \frac{1}{n}\sum_{\nu=1}^{n-M}\left(\frac{n\lambda}{\lambda_{\nu n}+n\lambda}\right)^2 z_{\nu n}^2 \Big/ \left(\frac{1}{n}\sum_{\nu=1}^{n-M}\frac{n\lambda}{\lambda_{\nu n}+n\lambda}\right)^2 \tag{4.6.2}$$

where $z_n = (z_{1n}, \ldots, z_{n-M,n})' = \Gamma' y$, where $\Gamma = Q_2 U$ as in (4.4.2). Provided the $\lambda_{\nu n}$ are nonzero,

$$\lim_{\lambda\to 0} V(\lambda) = \frac{1}{n}\sum \frac{z_{\nu n}{}^2}{\lambda_{\nu n}^2} \Big/ \left(\frac{1}{n}\sum \frac{1}{\lambda_{\nu n}}\right)^2 > 0. \tag{4.6.3}$$

However, unless $c(\lambda)$ has a pole of order at least $1/\lambda^2$ as $\lambda \to 0$, then $C(\lambda)$ of (4.6.1) will be zero at zero, so that in practice the criterion is unsuitable. For n large, the $\lambda_{\nu n}$ may be very small, and the calculation of V or C in the obvious way may be unstable near zero; this fact has possibly masked the unsuitability of certain criteria of this form C in Monte Carlo studies.

4.7 More on other estimates.

When σ^2 is known, an unbiased risk estimate is available for λ. This type of estimate was suggested by Mallows (1973) in the regression case, and applied to spline smoothing by Craven and Wahba (1979) (see also Hudson (1974)). Recalling that $ET(\lambda) = (1/n)\|(I - A(\lambda))g\|^2 + (\sigma^2/n)\operatorname{Tr} A^2(\lambda)$, we let

$$\hat{T}(\lambda) = \frac{1}{n}\|(I - A(\lambda)y\|^2 - \frac{\sigma^2}{n}\operatorname{Tr}(I - A(\lambda))^2 + \frac{\sigma^2}{n}\operatorname{Tr} A^2(\lambda). \tag{4.7.1}$$

It is not hard to show that $ET(\lambda) = E\hat{T}(\lambda)$. The numerical experiments in Craven and Wahba show that the GCV estimate and the unbiased risk estimate

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

behave essentially the same, to the accuracy of the experiment, when the same σ^2 is used in (4.7.1) as in generating the experimental data. It is probably true that a fairly good estimate of σ^2 would be required in practice to make this method work well. Several authors have suggested the so-called discrepancy method: Choose λ so that

$$\frac{1}{n}\|(I - A(\lambda))y\|^2 = \sigma^2. \tag{4.7.2}$$

The left-hand side is a monotone nondecreasing function of λ, and if $(1/n)\|(I - A(\infty))y\|^2$ (= the residual sum of squares after regression on the null space of $\|P_1(\cdot)\|^2$) is at least as large as σ^2, there will be a unique λ satisfying (4.7.2). We claim that this is not a very good estimate of the minimizer of $T(\lambda)$. Wahba (1975) showed that if λ^* is the minimizer of $ET(\lambda)$, then

$$E\frac{1}{n}\|(I - A(\lambda^*))y\|^2 = k\sigma^2(1 + o(1))$$

where k is a factor less than one. The experimental results in Craven and Wahba are consistent with these results, the discrepancy estimate λ_{dis} of λ being naturally larger than λ^* with $T(\lambda_{\text{dis}})/T(\lambda_{\text{opt}}) >> T(\lambda_{\text{GCV}})/T(\lambda_{\text{opt}})$, λ_{opt} being the minimizer of $T(\lambda)$.

By analogy with regression, I have suggested that $\operatorname{Tr} A(\lambda)$ be called the degrees of freedom for signal when λ is used. (Note that $M \leq$ d.f. signal $\leq n$), and this suggests an estimate for σ^2, as

$$\hat{\sigma}^2 = \hat{\sigma}^2(\hat{\lambda}) = \frac{\|(I - A(\hat{\lambda}))y\|^2}{\operatorname{Tr}(I - A(\hat{\lambda}))}$$

where $\hat{\lambda}$ is the GCV estimate of λ. Good numerical results for $\hat{\sigma}^2$ were obtained in Wahba (1983) although no theoretical properties of this estimate were given. Other estimates for σ^2 have been proposed (see, for example Buja, Hastie, and Tibshirani (1989). Hall and Titterington (1987) have proposed estimating λ as the solution to

$$\frac{\|(I - A(\lambda))y\|^2}{\operatorname{Tr}(I - A(\lambda))} = \sigma^2$$

when σ^2 is known. It is not known how this estimate would compare, say, with the unbiased risk estimate.

4.8 The generalized maximum likelihood estimate of λ.

A maximum likelihood estimate of λ based on the Bayes model was suggested by Anderssen and Bloomfield (1974) in the case of a stationary time series, and by Wecker and Ansley (1983) in the smoothing spline case (see also Barry (1983)).

Beginning with the stochastic model, (1.5.8) gives

$$y \sim N(0, b(\eta TT' + \Sigma + n\lambda I)) \tag{4.8.1}$$

where $\eta = a/b$ and a, b, T, and λ are as in (1.5.9), and $\lambda = \sigma^2/nb$.

Let

$$\begin{pmatrix} z \\ \cdots \\ w \end{pmatrix} = \begin{pmatrix} Q_2' \\ \cdots \\ \frac{1}{\sqrt{\eta}} T' \end{pmatrix} y, \tag{4.8.2}$$

where $Q_2'T = 0$, as in (1.3.18). Then

$$z \sim N(0, b(Q_2'\Sigma Q_2 + n\lambda I)), \tag{4.8.3}$$

$$\lim_{\eta \to \infty} Ezw' = 0,$$

$$\lim_{\eta \to \infty} Eww' = b(T'T)(T'T).$$

It was argued in Wahba (1985e) that the maximum likelihood estimate of λ here should be based on (4.8.3) since the distribution of w is independent of λ. This estimate was called the GML estimate. A straightforward maximization of the likelihood of (4.8.3) with respect to b and λ gives the GML estimate of λ as the minimizer of

$$\begin{aligned} M(\lambda) &= \frac{z'(Q_2'\Sigma Q_2 + n\lambda I)^{-1} z}{[\det (Q_2'\Sigma Q_2 + n\lambda I)^{-1}]^{1/(n-M)}} \\ &= \frac{y'Q_2(Q_2'\Sigma Q_2 + n\lambda I)^{-1}Q_2'y}{[\det (Q_2'\Sigma Q_2 + n\lambda I)^{-1}]^{1/(n-M)}}. \end{aligned}$$

Multiplying the top and bottom of this expression by $n\lambda$ results in an expression that is readily compared with $V(\lambda)$, viz.

$$M(\lambda) = \frac{y'(I - A(\lambda))y}{[\det^+(I - A(\lambda))]^{1/(n-M)}} \tag{4.8.4}$$

where $\det^+$ is the product of the nonzero eigenvalues. We remark that Wecker and Ansley (1983) included the M components of w as unknown parameters in the likelihood function. After minimizing with respect to w, they got a (slightly) different equation for "the" maximum likelihood estimate of λ. (See O'Hagan (1976) for other examples of this phenomenon.) We also note that if either σ^2 or b were known then a different expression for the maximum likelihood estimate of λ would be obtained.

It is shown in Wahba (1985e) that if $\|P_1 f\|^2 > 0$ and $\sum_{\nu=1}^{n-M} (h_{\nu n}^2/n)/(\lambda_{\nu n}/n)^p$ is bounded as $n \to \infty$ and $\mu_1(\lambda)$ and $\mu_2(\lambda)$ are $O(1/n\lambda^{1/q})$ for some $p \in (1,2]$, $q > 1$ then $(d/d\lambda)M(\lambda) = 0$ for $\lambda = \lambda_{\mathrm{GML}} = O(1/n^{q/(q+1)})$, independent of p. Thus, asymptotically, λ_{GML} is smaller than $\lambda^* = O(1/n^{q/(pq+1)})$, and an easy calculation shows that $\lim_{n\to\infty}(ET(\lambda_{\mathrm{GML}})/ET(\lambda^*)) \uparrow \infty$. On the other hand, it is argued in Wahba (1985e) that, if f is a sample function from a stochastic process with $Ef(s)f(t) = R(s,t)$, then the minimizers of both $V(\lambda)$ and $M(\lambda)$ estimate σ^2/nb. Thus, it is inadvisable to use the maximum likelihood estimate of λ since it is not robust against deviations from the stochastic model.

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 09/05/14 to 128.104.46.206. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

4.9 Limits of GCV.

The theory justifying the use of GCV is an asymptotic one. Good results cannot be expected for very small sample sizes when there is not enough information in the data to separate signal from noise. To take an extreme example, imagine, say, $n = 5$ data points (y_i, x_i) with x_i on the line. For arbitrary scattered values of the y_i's, given no further information, a curve interpolating the points, or the least squares straight line regression to the points, could be equally reasonable, and $V(\lambda)$ may well have minima at both zero and infinity. However, if even the order of magnitude of σ^2 is known in an example like this, then one could likely decide between these two extremes. My own Monte Carlo studies with smooth "truth" and independent and identically distributed Gaussian noise have resulted in generally reliable estimates of λ for n upwards of 25 or 30. It is to be noted that even for larger n, say $n = 50$, in extreme Monte Carlo replications there may be a handful of unwarranted extreme estimates ($\hat{\lambda} = 0$ or $\hat{\lambda} = \infty$), say a few percent, while the remaining estimates are all reasonable and more or less clustered together. This effect has been noted in Wahba (1983) and Section 6.3. Generally, if only σ^2 is known to within an order of magnitude, the occasional extreme case can be readily identified. As n gets larger, this effect becomes weaker, although it still defies ordinary statistical intuition. Even with "nice" examples with $n = 200$, there may be an occasional (2 or 3 out of 1,000, say) outliers in an otherwise "pleasant" population of sample $\hat{\lambda}$'s. One imagines that the theoretical distribution of $\hat{\lambda}$ can have (small) mass points at $\lambda = 0$ and $\lambda = \infty$ for moderate n.

My experience with GCV is that it is fairly robust against nonhomogeneity of variances and non-Gaussian errors (see, e.g., Villalobos and Wahba (1987)), and appears to work well when the ϵ_i's are due to quantization (see, e.g., Shahrary and Anderson (1989)). Andrews (1991) has recently provided some favorable theoretical results for unequal variances. However, the method is quite likely to give unsatisfactory results if the errors are highly correlated. It has given poor results when used to smooth a sample cumulative distribution F_n, for example, where $F_n(x_i) - F(x_i)$ and $F_n(x_j) - F(x_j)$ are correlated (Nychka, 1983) whereas differencing the data (see, e.g., Nychka et al. (1984)) so that the ϵ_i's are nearly independent has given good results. In a recent thesis, Altman (1987) discusses GCV in the presence of correlated errors. Of course if the noise is highly correlated, it becomes harder to distinguish it from "signal" by any nonparametric method that does not "know" anything about the nature of the correlation.

Trouble can arise with GCV if one has "exact" data (i.e., $\sigma^2 = 0$) *and* some of the $\lambda_{\nu n}$ appearing in (4.6.3) are insufficiently distinguishable from machine zero even though (in theory) they are strictly positive. In this case the theoretically "right" λ is zero, but in practice the numerical calculations with $\lambda = 0$ or λ near machine 0 can cause numerical instabilities and an unsatisfactory solution. Behavior of the $\lambda_{\nu n}$ in some well-known problems is discussed later.