A Note on Generalized Cross Validation with Replicates

by

Chong Gu, Nancy Heckman
and
Grace Wahba

\footnote{Research supported by the NSERC of Canada under Grant A7969, AFOSR under Grant AFOSR 87-0171 and NASA under Contract NAG 5-316. This report also appears as University of British Columbia Statistics Dept. TR 89}
A note on Generalized Cross Validation with replicates

Chong Gu, Nancy Heckman, and Grace Wahba
University of British Columbia, University of British Columbia
and University of Wisconsin—Madison

March 11, 1990

Abstract

Generalized Cross Validation (GCV) is a popular method for choosing the smoothing parameter in generalized spline smoothing, when there are independent errors with common unknown variance. When data points are replicated, then one has an independent estimate of the unknown variance $\sigma^2$. One may then ask how best to use this information. For example, one may use the estimate of $\sigma^2$ in an unbiased risk estimate for the smoothing parameter, instead of using GCV. In this note we show, that as the number of degrees of freedom for the estimate of $\sigma^2$ tends to infinity, the GCV estimate and the unbiased risk estimate of Craven and Wahba become identical.

Key words and phrases: generalized cross validation, unbiased risk estimate.

1 Introduction

We first introduce notations and results for calculating the generalized cross validation (hereafter GCV) score and an unbiased estimate of risk in a general setting. These results will be applied to the particular case of replicated data points in Section 2.

Suppose one observes \( y_j = f(x_j) + \epsilon_j, j = 1, \ldots, n \), where \( x_j \in X \), \( E\epsilon_j = 0, \) \( \text{Var}(\epsilon_j) = w_j^{-1}\sigma^2 \) with \( \sigma^2 \) unknown, and the \( \epsilon_j \)'s uncorrelated. A smoothing spline is the solution \( f_\lambda \) to the penalized least squares problem

\[
\min \sum_{j=1}^{n} w_j(y_j - f(x_j))^2 + \lambda \| P_1 f \|^2, \quad \text{s.t. } f \in \mathcal{H}
\]

(1)

where \( \mathcal{H} \) is a reproducing kernel Hilbert space of functions on the domain \( X \) with norm \( \| \cdot \| \), and \( P_1 \) is the orthogonal projector onto a subspace \( H_1 \) of co-dimension \( M < n \). \( f_\lambda \) has an expression \( \sum_{i=1}^{M} d_i \phi_i(\cdot) + \sum_{j=1}^{n} c_j R_1(x_j, \cdot) \), where \( \{ \phi_i \}_{i=1}^{M} \) is a basis for \( H_0 \), the orthogonal complement of \( H_1 \), and \( R_1 \) is the reproducing kernel of \( H_1 \). Substituting the solution expression into (1), one solves

\[
\min (y - Qc - Sd)^TW(y - Qc - Sd) + \lambda e^TQCe
\]

(2)

for \( c \) and \( d \), where \( (Q)_{j,k} = R_1(x_j, x_k), (S)_{j,\nu} = \phi_\nu(x_j), \) and \( W = \text{diag}(w_1, \ldots, w_n) \). See Wahba (1990).

For the standard setting where \( w_j = 1 \) and \( W = I \), defining the hat matrix \( A(\lambda) \) satisfying \( \hat{y} = Qc + Sd = A(\lambda)y \), Craven and Wahba (1979) proposed choosing the smoothing parameter \( \lambda \) as the minimizer of the GCV score

\[
V(\lambda) = \frac{(y - \hat{y})^TW(y - \hat{y})}{\text{tr}(I - A(\lambda))^2},
\]

(3)

and they argued that the GCV method is asymptotically optimal for minimizing the expected predictive mean square error. Stronger optimality results are found in Li (1986) and references cited there. For \( W \neq I \), the natural extension of the GCV score is

\[
V_W(\lambda) = \frac{(y - \hat{y})^TW(y - \hat{y})}{\text{tr}(I - A_W(\lambda))^2},
\]

(4)
where $A_W$ satisfies $W^{1/2}y = A_W(W^{1/2}y)$. $V_W$ has a similar asymptotic optimality as that of $V$; see O’Sullivan et al. (1986), Section 3.1. In general, it can be shown that

$$I - A_W = \lambda F_2(F_2^T W^{1/2}QW^{1/2}F_2 + \lambda I)^{-1}F_2^T$$

(5)

where $F_2^T F_2 = I_{n-M}$ and $F_2^T W^{1/2} S = 0$. See, e.g., Gu et al. (1989), for results when $W = I$.

Now suppose $\sigma^2$ is known. Letting $f$ denote the true function evaluated at the data points $x_i$, it is easily shown that $E(y - \hat{y})^T W (y - \hat{y}) = (f - W^{-1/2}A_W W^{1/2}f)^T W (f - W^{-1/2}A_W W^{1/2}f) + \sigma^2 \text{tr}(I - A_W(\lambda))^2$ and the risk (weighted mean square error) is $E(f - \hat{y})^T W (f - \hat{y}) = (f - W^{-1/2}A_W W^{1/2}f)^T W (f - W^{-1/2}A_W W^{1/2}f) + \sigma^2 \text{tr} A_W(\lambda)^2$. Based on this, Craven and Wahba (1979) proposed a method of choosing $\lambda$ for known $\sigma^2$ by minimizing an unbiased estimate of the risk, namely

$$(y - \hat{y})^T W (y - \hat{y}) - \sigma^2 \text{tr}(I - 2A_W(\lambda)).$$

(6)

Another approach in the $\sigma^2$ known case was proposed by Hall and Titterington (1987). They suggested that $\lambda$ be chosen to satisfy

$$\frac{(y - \hat{y})^T W (y - \hat{y})}{\text{tr}(I - A_W(\lambda))} = \sigma^2$$

(7)

since the left hand side of (7) is believed to behave well as an estimate of $\sigma^2$.

Douglas Bates (personal communication), Dolph Schluter (personal communication), and others have asked what one should do to estimate $\lambda$ if, in the $\sigma^2$ unknown case, the data points $x_j$ are not distinct. In that case, one could obtain (the usual) independent estimate $\hat{\sigma}^2$ of $\sigma^2$. One possibility is simply to ignore this fact and minimize (4). It is not hard to see that the various optimality properties of the GCV estimate are not lost. Another possibility is to substitute the estimate $\hat{\sigma}^2$ for $\sigma^2$ in either (6) or (7). We do not further discuss the latter option. Our results are on the former option. In this note, we show that, in the limit as the number of degrees of freedom for $\hat{\sigma}^2$ tends to infinity, the GCV estimate obtained by minimizing (4) is equal to the unbiased risk estimate obtained by minimizing (6) with $\sigma^2$ replaced by $\hat{\sigma}^2$. 


2 Results

Consider
\[ y_{jk} = f(x_j) + \epsilon_{jk}, \quad j = 1, \ldots, n; k = 1, \ldots, r_j \]
where the \( \epsilon_{jk} \)'s are uncorrelated with mean 0 and unknown variance \( \sigma^2 \). Let \( N = \sum_{j=1}^{r_j} r_j \) and define \( P \) to be the \( N \times n \) matrix \( \text{diag}(1_{r_j}) \). The pooled data vector is \( \tilde{y} = W^{-1} P^T y \), where \( y = (y_{1,1}, \ldots, y_{1,r_1}, \ldots, y_{n,1}, \ldots, y_{n,r_n})^T \) and \( W = (P^T P) = \text{diag}(r_1, \ldots, r_n) \). All information about \( f \) is contained in \( \tilde{y} \). The GCV score for the pooled data is
\[
\hat{V}(\lambda) = \frac{(\tilde{y} - \hat{y})^T W (\tilde{y} - \hat{y})}{\text{tr}(I - \tilde{A}_W(\lambda))^2},
\]
where \( I - \tilde{A}_W = \lambda \tilde{F}_2 (\tilde{F}_2 W^{1/2} \tilde{Q} W^{1/2} \tilde{F}_2 + \lambda I)^{-1} \tilde{F}_2^T \), and the bars indicate quantities associated with the pooled data. However the pooled data GCV score has lost the information contained in the scatter of the replicates about their means.

Now we return to the original data \( y \) and reexamine (4) with \( W = I_N \). By (5) we need to find \( F_2 \) such that \( F_2^T F_2 = I_{N-M} \) and \( F_2^T S = 0 \). There exists \( F_3 \) such that \( F_3^T F_3 = I_{N-n} \) and \( F_3^T P = 0 \). It is easy to verify that \( F_2 = (PW^{-1/2} \tilde{F}_2 : F_3) \) is orthogonal and \( F_2^T S = F_2^T P \).

Following this, it can be shown that \( y^T (I - A)^2 y = (\tilde{y} - \hat{y})^T W (\tilde{y} - \hat{y}) + (N - n) \hat{\sigma}^2 \), where \( \hat{\sigma}^2 = y^T F_3 F_2^T y / (N - n) \) is the (usual) unbiased estimate of \( \sigma^2 \) with \( N - n \) degrees of freedom. Similarly, \( \text{tr}(I - A) = \text{tr}(I_N - A_W) + (N - n) \). So the full data GCV score (which is equal to (4)) is
\[
V(\lambda) = \frac{(\tilde{y} - \hat{y})^T W (\tilde{y} - \hat{y}) + (N - n) \hat{\sigma}^2}{\text{tr}(I - \tilde{A}_W(\lambda)) + (N - n)^2}.
\]
(8)
Furthermore the unbiased risk estimate of \( \lambda \) is the minimizer of
\[
(\tilde{y} - \hat{y})^T W (\tilde{y} - \hat{y}) - \sigma^2 \text{tr}(I - 2 \tilde{A}_W),
\]
(9)
since (9) differs from (6) by a quantity which does not depend on \( \lambda \). The minimizer of (9) satisfies
\[
\frac{d}{d\lambda}(\tilde{y} - \hat{y})^T W (\tilde{y} - \hat{y}) = -2 \hat{\sigma}^2 \frac{d}{d\lambda} \text{tr}\tilde{A}_W,
\]
(10)
and the minimizer of (8) satisfies

$$\frac{d}{d\lambda}(\hat{y} - \hat{\hat{y}})^TW(\hat{y} - \hat{\hat{y}}) = -2\frac{d}{d\lambda} \frac{\text{tr}(\hat{A}W)(\hat{y} - \hat{\hat{y}})^TW(\hat{y} - \hat{\hat{y}}) + (N - n)\hat{\sigma}^2}{\text{tr}(I - \hat{A}W) + (N - n)}.$$  (11)

As $(N - n) \to \infty$, $\hat{\sigma}^2 \overset{a.s.}{\to} \sigma^2$ and $[(\hat{y} - \hat{\hat{y}})^TW(\hat{y} - \hat{\hat{y}}) + (N - n)\hat{\sigma}^2]/[\text{tr}(I - \hat{A}W) + (N - n)] \overset{a.s.}{\to} \sigma^2$, giving the claimed result.

**Acknowledgements**

This research was supported by the NSERC of Canada under Grant A-7969, by the AFOSR under Grant AFOSR 87-0171 and by NASA under Contract NAG5-316.

**References**


