DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 864

March 1990

A Note on Generalized Cross Validation with Replicates [1]

by

Chong Gu, Nancy Heckman

and

Grace Wahba

[1]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 σ^2. One may then ask how best to use this information. For example, one may use the estimate of σ^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 σ^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.*

AMS 1980 subject classifications. Primary 65D07, 65D10.

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 $y_j = f(\boldsymbol{x}_j) + \epsilon_j$, $j = 1, \cdots, n$, where $\boldsymbol{x}_j \in \mathcal{X}$, $E\epsilon_j = 0$, $Var(\epsilon_j) = w_j^{-1}\sigma^2$ with σ^2 unknown, and the ϵ_j's uncorrelated. A smoothing spline is the solution f_λ to the penalized least squares problem

$$\min \sum_{j=1}^{n} w_j (y_j - f(\boldsymbol{x}_j))^2 + \lambda \|P_1 f\|^2, \qquad s.t.\ f \in \mathcal{H} \tag{1}$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space of functions on the domain $\mathcal{X}$ with norm $\|\cdot\|$, and P_1 is the orthogonal projector onto a subspace $\mathcal{H}_1$ of co-dimension $M < n$. f_λ has an expression $\sum_{\nu=1}^{M} d_\nu \phi_\nu(\cdot) + \sum_{j=1}^{n} c_j R_1(\boldsymbol{x}_j, \cdot)$, where $\{\phi_\nu\}_{\nu=1}^{M}$ is a basis for $\mathcal{H}_0$, the orthogonal complement of $\mathcal{H}_1$, and R_1 is the reproducing kernel of $\mathcal{H}_1$. Substituting the solution expression into (1), one solves

$$\min\, (\boldsymbol{y} - Q\boldsymbol{c} - S\boldsymbol{d})^T W (\boldsymbol{y} - Q\boldsymbol{c} - S\boldsymbol{d}) + \lambda \boldsymbol{c}^T Q \boldsymbol{c} \tag{2}$$

for $\boldsymbol{c}$ and $\boldsymbol{d}$, where $(Q)_{j,k} = R_1(\boldsymbol{x}_j, \boldsymbol{x}_k)$, $(S)_{j,\nu} = \phi_\nu(\boldsymbol{x}_j)$, and $W = \mathrm{diag}(w_1, \cdots, w_n)$. See Wahba (1990).

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

$$V(\lambda) = \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}})}{[\mathrm{tr}(I - A(\lambda))]^2}, \tag{3}$$

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

$$V_W(\lambda) = \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}})^T W (\boldsymbol{y} - \hat{\boldsymbol{y}})}{[\mathrm{tr}(I - A_W(\lambda))]^2}, \tag{4}$$

where A_W satisfies $W^{1/2}\hat{\boldsymbol{y}} = A_W(W^{1/2}\boldsymbol{y})$. 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

$$I - A_W = \lambda F_2(F_2^T W^{1/2} Q W^{1/2} F_2 + \lambda I)^{-1} F_2^T \tag{5}$$

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

$$(\boldsymbol{y}-\hat{\boldsymbol{y}})^T W(\boldsymbol{y}-\hat{\boldsymbol{y}}) - \sigma^2 \mathrm{tr}(I - 2A_W(\lambda)). \tag{6}$$

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

$$\frac{(\boldsymbol{y}-\hat{\boldsymbol{y}})^T W(\boldsymbol{y}-\hat{\boldsymbol{y}})}{\mathrm{tr}(I - A_W(\lambda))} = \sigma^2 \tag{7}$$

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

Douglas Bates (personal communication), Dolph Schluter (personal communication), and others have asked what one should do to estimate λ if, in the σ^2 unknown case, the data points $\boldsymbol{x}_j$ are not distinct. In that case, one could obtain (the usual) independent estimate $\hat{\sigma}^2$ of σ^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 σ^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 σ^2 replaced by $\hat{\sigma}^2$.

2 Results

Consider

$$y_{jk_j} = f(\boldsymbol{x}_j) + \epsilon_{jk_j}, \qquad j = 1, \cdots, n; k_j = 1, \cdots, r_j$$

where the ϵ_{jk_j}'s are uncorrelated with mean 0 and unknown variance σ^2. Let $N = \sum_{j=1}^{n} r_j$ and define P to be the $N \times n$ matrix $\text{diag}(\mathbf{1}_{r_j})$. The pooled data vector is $\bar{\boldsymbol{y}} = W^{-1}P^T\boldsymbol{y}$, where $\boldsymbol{y} = (y_{1,1}, \cdots, y_{1,r_1}, \cdots, y_{n,1}, \cdots, y_{n,r_n})^T$ and $W = (P^TP) = \text{diag}(r_1, \cdots, r_n)$. All information about f is contained in $\bar{\boldsymbol{y}}$. The GCV score for the pooled data is

$$\bar{V}(\lambda) = \frac{(\bar{\boldsymbol{y}} - \hat{\bar{\boldsymbol{y}}})^T W (\bar{\boldsymbol{y}} - \hat{\bar{\boldsymbol{y}}})}{[\text{tr}(I - \bar{A}_W(\lambda))]^2},$$

where $I - \bar{A}_W = \lambda \bar{F}_2 (\bar{F}_2 W^{1/2} \bar{Q} W^{1/2} \bar{F}_2 + \lambda I)^{-1} \bar{F}_2^T$, 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 $\boldsymbol{y}$ 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 $F_2 = (PW^{-1/2}\bar{F}_2 : F_3)$ is orthogonal and $F_2^T S = F_2^T P \bar{S} = 0$. Following this, it can be shown that $\boldsymbol{y}^T(I - A)^2\boldsymbol{y} = (\bar{\boldsymbol{y}} - \hat{\bar{\boldsymbol{y}}})^T W (\bar{\boldsymbol{y}} - \hat{\bar{\boldsymbol{y}}}) + (N - n)\hat{\sigma}^2$, where $\hat{\sigma}^2 = \boldsymbol{y}^T F_3 F_3^T \boldsymbol{y}/(N - n)$ is (the usual) unbiased estimate of σ^2 with $N - n$ degrees of freedom. Similarly, $\text{tr}(I - A) = \text{tr}(I_n - \bar{A}_W) + (N - n)$. So the full data GCV score (which is equal to (4))is

$$V(\lambda) = \frac{(\bar{\boldsymbol{y}} - \hat{\bar{\boldsymbol{y}}})^T W (\bar{\boldsymbol{y}} - \hat{\bar{\boldsymbol{y}}}) + (N - n)\hat{\sigma}^2}{[\text{tr}(I - \bar{A}_W(\lambda)) + (N - n)]^2}. \tag{8}$$

Furthermore the unbiased risk estimate of λ is the minimizer of

$$(\bar{\boldsymbol{y}} - \hat{\bar{\boldsymbol{y}}})^T W (\bar{\boldsymbol{y}} - \hat{\bar{\boldsymbol{y}}}) - \hat{\sigma}^2 \text{tr}(I - 2\bar{A}_W), \tag{9}$$

since (9) differs from (6) by a quantity which does not depend on λ. The minimizer of (9) satisfies

$$\frac{d}{d\lambda}(\bar{\boldsymbol{y}} - \hat{\bar{\boldsymbol{y}}})^T W (\bar{\boldsymbol{y}} - \hat{\bar{\boldsymbol{y}}}) = -2\hat{\sigma}^2 \frac{d}{d\lambda} \text{tr}\bar{A}_W, \tag{10}$$

and the minimizer of (8) satisfies

$$\frac{d}{d\lambda}(\bar{y}-\hat{\bar{y}})^T W(\bar{y}-\hat{\bar{y}}) = -2\left(\frac{d}{d\lambda}\mathrm{tr}\bar{A}_W\right)\frac{(\bar{y}-\hat{\bar{y}})^T W(\bar{y}-\hat{\bar{y}}) + (N-n)\hat{\sigma}^2}{\mathrm{tr}(I-\bar{A}_W)+(N-n)}. \tag{11}$$

As $(N-n)\to\infty$, $\hat{\sigma}^2 \overset{a.s.}{\to} \sigma^2$ and $[(\bar{y}-\hat{\bar{y}})^T W(\bar{y}-\hat{\bar{y}}) + (N-n)\hat{\sigma}^2]/[\mathrm{tr}(I-\bar{A}_W)+(N-n)] \overset{a.s.}{\to} \sigma^2$, 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

Craven, P. and Wahba, G. (1979), "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerische Mathematik*, 31, 377 – 403.

Gu, C., Bates, D.M., Chen, Z., and Wahba, G. (1989), "The computation of GCV functions through householder tridiagonalization with application to the fitting of interaction spline models," *SIAM Journal on Matrix Analysis and Applications*, 10, 457 – 480.

Hall, P. and Titterington, D. (1987), "Common structure of techniques for choosing smoothing parameters in regression problems, " *Journal of the Royal Statistical Society Ser. B*, 49, 184 – 198.

Li, K. C. (1986), " Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing", *Annals of Statistics*, 14, 1101 – 1112.

O'Sullivan, F., Yandell, B., and Raynor, W. (1986), "Automatic smoothing of regression functions in generalized linear models," *Journal of the American Statistical Association*, 81, 96 – 103.

Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, 59, SIAM.