

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 906

Note: Cowles Foundation Discussion Papers are preliminary materials circulated to stimulate discussion and critical comment. Requests for single copies of a Paper will be filled by the Cowles Foundation within the limits of the supply. References in publications to Discussion Papers (other than acknowledgment that a writer had access to such unpublished material) should be cleared with the author to protect the tentative character of these papers.

ASYMPTOTIC OPTIMALITY OF GENERALIZED C_L , CROSS-VALIDATION,
AND GENERALIZED CROSS-VALIDATION IN REGRESSION WITH
HETEROSKEDASTIC ERRORS

by

Donald W.K. Andrews

May 1988

Revised: May 1989

ASYMPTOTIC OPTIMALITY OF GENERALIZED C_L ,
CROSS-VALIDATION, AND GENERALIZED CROSS-VALIDATION
IN REGRESSION WITH HETEROSKEDASTIC ERRORS

Donald W. K. Andrews*

Cowles Foundation
Yale University

May, 1988
Revised: May, 1989

ABSTRACT

The problem considered here is that of using a data-driven procedure to select a good estimate from a class of linear estimates indexed by a discrete parameter. In contrast to other papers on this subject, we consider models with heteroskedastic errors. The results apply to model selection problems in linear regression and to nonparametric regression estimation via series estimators, nearest neighbor estimators, and local regression estimators, among others. Generalized C_L (GC_L), cross-validation (CV), and generalized cross-validation (GCV) procedures are analyzed. The GC_L and CV criteria are shown to be asymptotically optimal under general conditions. The GCV criterion is found to be asymptotically optimal only under a condition that is satisfied in some applications but not in others. For example, it is satisfied in the nearest neighbor estimation context but not in the series estimation, local regression estimation, or model selection contexts. The proofs rely heavily on results of Li (1987).

JEL classification number: 211.

Key words: Additive interactive regression model, cross-validation, generalized C_L , generalized cross-validation, heteroskedastic errors, interactive spline estimators, local regression estimator, model selection, nearest neighbor estimators, nonparametric regression, ridge regression estimators, series estimators, spline estimators.

1. Introduction

Suppose the observations $y_n = (y_1, \dots, y_n)'$ satisfy the model

$$y_i = \mu_i + e_i, \quad i = 1, 2, \dots, n,$$

where $\underline{\mu}_n = (\mu_1, \dots, \mu_n)'$ is the unobserved mean vector of y_n and $e_n = (e_1, \dots, e_n)'$ is the unobserved error vector comprised of independent, mean zero, variance σ_i^2 errors. Consider a class of linear estimators $\hat{\underline{\mu}}_n(h) = M_n(h)y_n$, where each estimator is indexed by a parameter h in an index set H_n . Here $M_n(h)$ is an $n \times n$ nonrandom matrix that may depend on some nonrandom regressor variables as well as the parameter h . The object is to use the observed vector y_n to select \hat{h} from H_n in such a way as to make the average squared prediction error

$$L_n(\hat{h}) = n^{-1} \|y_n - M_n(\hat{h})y_n\|^2$$

as small as possible (where $\|\cdot\|$ denotes the Euclidean norm).

This problem has been analyzed by Li (1987) and others for the case where the error variances σ_i^2 are homoskedastic. Here we extend the results of Li (1987) to the case of heteroskedastic errors. For ease of comparison, we adopt the same notation and numbering of assumptions and equations as in Li (1987). Assumptions and equations that appear in this paper but not in Li's are denoted by asterisks; those without asterisks are the same as in Li's.

Examples of the problem outlined above include:

EXAMPLE 1. Model Selection: Associated with each y_i there are p_n explanatory variables x_{i1}, \dots, x_{ip_n} arranged in decreasing order of importance. A linear model $\mu_i = \sum_{j=1}^h x_{ij} \beta_j$ is proposed based on the first h variables and one estimates $\underline{\mu}_n$ using the least squares estimator $\hat{\underline{\mu}}_n(h) = X_h(X_h'X_h)^{-1}X_h'y_n$. Here X_h is the (full rank) $n \times h$ design matrix with (i,j) -th element x_{ij} and $M_n(h)$ is the projection matrix $X_h(X_h'X_h)^{-1}X_h'$. The index set H_n is $(1, \dots, p_n)$. The goal is to determine an appropriate model for the purpose of estimating $\underline{\mu}_n$.

EXAMPLE 2. Series Estimation of a Nonparametric Regression Model [Gallant (1981), Geman and Hwang (1982), Andrews (1988)]: The mean of y_i is an unknown function $f(\cdot)$ of an observed regressor vector z_i in $Z \subset R^p$, $\mu_i = f(z_i)$. A series approximation of $f(z_i)$ is constructed based on h terms, $\sum_{j=1}^h x_j(z_i) \beta_j$, where the functions $x_j(\cdot)$, $j = 1, \dots, h$ are known (e.g., trigonometric functions) and the coefficients β_j , $j = 1, \dots, h$ are unknown. One estimates $\underline{\mu}_n$ by estimating the unknown constants $\{\beta_j\}$ by least squares: $\hat{\underline{\mu}}_n(h) = X_h(X_h'X_h)^{-1}X_h'y_n$, where X_h is the (full rank) $n \times h$ matrix with (i,j) -th element $x_j(z_i)$. As in Example 1, $M_n(h) = X_h(X_h'X_h)^{-1}X_h'$. The index set H_n is $(1, \dots, n)$ or some subset thereof. The goal is to determine the appropriate number of terms in the series expansion to be used in estimating $\underline{\mu}_n$.

EXAMPLE 3. Series Estimation of an Additive Interactive Regression (AIR) Model [Andrews (1988), Andrews and Whang (1989)]: The model is the same as in Example 2 except that $f(\cdot)$ is known to be of the form $f(\cdot) = \sum_{a=1}^A \sum_{b=1}^{B(a)} f_{ab}(\cdot)$ for unknown functions $\{f_{ab}(\cdot)\}$, where $f_{ab}(z_i)$ depends on only "a" ($\leq d$) different elements of the d -vector z_i for each

$b = 1, \dots, B(a)$. For example, one might have $f_{1b}(z_i) = f_{1b}^*(z_{1b})$ and $f_{2b}(z_i) = f_{2b}^*(z_{i1}, z_{i2})$, where $z_i = (z_{i1}, \dots, z_{id})'$. If $A = 1$, the model is an additive regression model. If $A > 1$, the model allows interactions between the elements of z_i . If $A = d$ and all the interaction terms are included, then the model is a fully nonparametric regression model. As shown in Andrews and Whang (1989), the rate of convergence of series estimators in AIR models depends on A and not on d , and hence, the curse of dimensionality is circumvented. Typically A is taken to be quite small, e.g., one or two, and some of the possible interaction functions $f_{ab}(\cdot)$ for each $a = 1, \dots, A$ are excluded.

A series approximation of $f(z_i)$ is constructed using a series approximation $\sum_{c=1}^{h_{ab}} x_{abc}(z_i)\beta_{abc}$ of each function $f_{ab}(z_i)$, where $\{\beta_{abc}\}$ are unknown coefficients and $\{x_{abc}(\cdot)\}$ are known functions (e.g., trigonometric functions) that depend on the same elements of z_i as does $f_{ab}(\cdot)$ for all $c = 1, \dots, h_{ab}$. One estimates μ_n by using least squares to estimate $\{\beta_{abc}\}$: $\hat{\mu}_n(h) = X_h'(X_h'X_h)^{-1}X_h'y_n$, where X_h is the (full rank) $n \times (\sum_{a=1}^A \sum_{b=1}^{B(a)} h_{ab})$ matrix with i -th row given by the elements of $\{x_{abc}(z_i) : c = 1, \dots, h_{ab}; b = 1, \dots, B(a); a = 1, \dots, A\}$. The parameter h in this example is a vector $(h_{11}, \dots, h_{1B(1)}, h_{21}, \dots, h_{AB(A)})'$ of non-negative integers of dimension $D = \sum_{a=1}^A B(a)$. The index set H_n is some subset of $\{h \in I_+^D : h' \underline{1} \leq n\}$, where I_+ denotes the set of non-negative integers and $\underline{1}$ denotes a D -vector of ones. The goal is to determine the appropriate number of terms h_{ab} in the series expansion of each of the functions $f_{ab}(\cdot)$.

EXAMPLE 4. Nearest-neighbor Estimation of a Nonparametric Regression Model [Stone (1977)]: The model is as in Example 2. Let $z_{i(j)}$ denote the j -th nearest neighbor of z_i in the sense that $\|z_i - z_{i(j)}\|$ is the j -th smallest number among the n values $\|z_i - z_v\|$, $v = 1, \dots, n$. (Ties may be broken in any systematic fashion.) For a given weight function $w_{n,h}(\cdot)$, the h -nearest-neighbor estimate of μ_i is $\hat{\mu}_i(h) = \sum_{j=1}^h w_{n,h}(j)y_{i(j)}$. Hence, $\hat{\mu}_n(h)$ is of the form $M_n(h)y_n$, where each row of $M_n(h)$ is some permutation of the n -vector $(w_{n,h}(1), \dots, w_{n,h}(h), 0, \dots, 0)$. Uniform, triangular, and quadratic weights among others have been considered in the literature (see Stone (1977, p. 600)). Assumptions on the weights $w_{n,h}(\cdot)$ are specified below. The index set H_n is $\{1, \dots, n\}$ or some subset thereof. The goal is to use y_n to determine the number of neighbors to include in the estimate of μ_n .

Additional examples include local regression estimation of nonparametric regression models (see Cleveland and Devlin (1988)), kernel nonparametric regression estimation with single or multiple smoothing parameters (e.g., see Bierens (1987)), smoothing spline nonparametric regression estimation (see Wahba (1989)), interaction spline nonparametric regression estimation with multiple smoothing parameters (see Wahba (1986)), and ridge regression estimation. The latter four examples apply only if one restricts (somewhat unnaturally) the possible values of the smoothing parameter to a finite grid of points that increases with the sample size.

We analyze three different procedures for selecting h :

(i) Generalized C_L (GC_L): Select \hat{h} , denoted by \hat{h}_M , that achieves

$$\min_{h \in H_n} n^{-1} \|y_n - \hat{\mu}_n(h)\|^2 + 2n^{-1} \text{tr } M_n(h)\Omega, \quad (1.1^*)$$

where Ω is an $n \times n$ diagonal matrix with diagonal elements $\{\sigma_1^2, \dots, \sigma_n^2\}$. This procedure is a generalization for models with heteroskedastic errors of Mallows' (1973) C_L procedure. In the model selection example, GC_L is a generalization of Mallows' well-known C_p procedure.

(ii) Generalized cross-validation (GCV) [Craven and Wahba (1979)]: Select \hat{h} , denoted by \hat{h}_G , that achieves

$$\min_{h \in H_n} \frac{n^{-1} \|\mathbf{y}_n - \hat{\mu}_n(h)\|^2}{(1 - n^{-1} \text{tr } M_n(h))^2}, \quad (1.2)$$

where $\text{tr } M_n(h)$ denotes the trace of the matrix $M_n(h)$.

(iii) (Delete one) Cross-validation [Allen (1974), Stone (1974), Geisser (1975), Wahba and Wold (1975)]: Select \hat{h} , denoted by \hat{h}_c , that minimizes the sum of squared prediction errors for y_i , where the predictor of y_i is based on the estimator of μ_n that uses all of the data except y_i . The form of this predictor depends on the definition of $\hat{\mu}_n(h)$ when the sample size is $n-1$. Given $y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n$ write the predictor of y_i as $\hat{y}_{-i} = \sum_{j=1}^n \tilde{m}_{ij}(h) y_j$ with $\tilde{m}_{ii}(h) = 0$. Then \hat{h}_c achieves

$$\min_{h \in H_n} n^{-1} \|\mathbf{y}_n - \tilde{M}_n(h) \mathbf{y}_n\|^2, \quad (1.3)$$

where $\tilde{M}_n(h)$ is the $n \times n$ matrix with $\tilde{m}_{ij}(h)$ as the (i,j) -th entry.

Note that the GC_L procedure requires knowledge of the error variances $\{\sigma_1^2, \dots, \sigma_n^2\}$ whereas the GCV and CV procedures do not. In Example 4, however, a "feasible" analogue of GC_L , which does not require knowledge of the error variances, can be considered (see Section 2 below).

In the examples considered above, the selection procedures simplify.

In Examples 1-3,

$$\tilde{M}_n(h) = D_n(h)(M_n(h) - I_n) + I_n, \quad (1.4)$$

where $D_n(h)$ is an $n \times n$ diagonal matrix with i -th diagonal element equal to $(1 - m_i(h))^{-1}$, $m_i(h)$ is the i -th diagonal element of the matrix $M_n(h)$, and I_n is the n -dimensional identity matrix (see Li (1987, p. 960)). In this case, the CV criterion (1.3) becomes

$$\min_{h \in H_n} n^{-1} \sum_{i=1}^n (y_i - \hat{\mu}_i(h))^2 / (1 - m_i(h))^2. \quad (1.5)$$

In Examples 1-4, $n^{-1} \text{tr } M_n(h)$ equals h/n , h/n , $h'1/n$, and $w_{n,h}(1)$ respectively. In Example 4, $n^{-1} \text{tr } M_n(h)\Omega = w_{n,h}(1)n^{-1} \sum_{i=1}^n \sigma_i^2$, $\hat{y}_{-i} = \sum_{j=1}^h w_{n,h}(j)y_{i(j+1)}$, and $\tilde{M}_n(h)$ has rows that are permutations of $(w_{n,h}(1), \dots, w_{n,h}(h), 0, \dots, 0)$ and diagonal elements that are zeroes.

We are interested in determining conditions under which the above procedures are *asymptotically optimal* in the sense that

$$\frac{\hat{L}_n(h)}{\inf_{h \in H_n} \hat{L}_n(h)} \xrightarrow{P} 1 \text{ and} \quad (1.6)$$

$$\frac{\hat{R}_n(h)}{\inf_{h \in H_n} \hat{R}_n(h)} \rightarrow 1 \quad (1.7^*)$$

even when the errors are heteroskedastic, where $R_n(h) = EL_n(h)$ and " $\xrightarrow{P} 1$ " denotes convergence in probability as $n \rightarrow \infty$.

In Section 2 we find that analogues of Li's (1987) conditions for the asymptotic optimality of Mallows's C_L procedure when the errors are homoske-

dastic can be used to establish the optimality of GC_L when the errors are heteroskedastic. In Section 3, we use Li's notion of a nil-trace estimator to obtain the asymptotic optimality of GCV from the asymptotic optimality of GC_L . To do so, a condition is needed that is satisfied in Example 4 but not in Examples 1-3. This condition is satisfied if the diagonal elements of $M_n(h)$ are all equal. In Section 4, the asymptotic optimality of CV is established using the results of Section 2 for GC_L . The conditions used here are analogous to those of Li for the homoskedastic error case. In particular, the additional condition used in the treatment of GCV is not needed and CV is asymptotically optimal in all of the examples. Section 5 contains proofs of the results.

2. Generalized C_L

We have

$$R_n(h) = EL_n(h) = n^{-1} \|A_n(h)\underline{\mu}_n\|^2 + n^{-1} \text{tr } M'_n(h)M_n(h)\Omega ,$$

where $A_n(h) = I_n - M_n(h)$. Let $\lambda(M_n(h))$ denote the largest eigenvalue of $M_n(h)$. The asymptotic optimality of GC_L is established under the assumptions

$$\overline{\lim}_{n \rightarrow \infty} \sup_{h \in H_n} \lambda(M_n(h)) < \infty , \quad (\text{A.1})$$

$$\sup_{i \geq 1} Ee_i^{4m} < \infty , \quad 0 < \inf_{i \geq 1} \sigma_i^2 \leq \sup_{i \geq 1} \sigma_i^2 < \infty , \quad (\text{A.2*})$$

$$\sum_{h \in H_n} (nR_n(h))^{-m} \rightarrow 0 \text{ as } n \rightarrow \infty , \quad (\text{A.3})$$

for some positive integer m . (Assumption (A.1) of Li (1987, p. 961) has "lim" in place of " $\overline{\lim}$," but the latter undoubtedly was intended.)

Under assumption (A.2*), assumption (A.3) holds if and only if it holds with Ω replaced by $\sigma^2 I_n$ for arbitrary σ^2 in $[\inf_{i \geq 1} \sigma_i^2, \sup_{i \geq 1} \sigma_i^2]$. Thus, (A.3) in the case of heteroskedastic errors is no stronger than it is in the case of homoskedastic errors.

THEOREM 2.1*. Under assumptions (A.1), (A.2*), and (A.3), GC_L is asymptotically optimal, i.e., (1.6) and (1.7*) hold with $\hat{h} = \hat{h}_M$.

EXAMPLES 1-3 (cont.). In these examples, (A.1) is automatically satisfied, since $M_n(h)$ is a projection matrix. In addition, in Examples 1 and 2, (A.3) with $m = 2$ can be replaced by

$$\inf_{h \in H_n} nR_n(h) \rightarrow \infty, \quad (\text{A.3}')$$

since (A.2*) and (A.3') imply (A.3) with $m = 2$. To see the latter, apply Li's (1987) argument given in his equations (2.5) and (2.6) with $h\sigma^2$ replaced by $\text{tr } M_n(h)\Omega$ and $h \inf_{i \geq 1} \sigma_i^2$ in the two places it appears in (2.5) and replace σ^{-4} by $(\inf_{i \geq 1} \sigma_i^2)^{-2}$ where it appears in (2.6). In Example 3, (A.3) with $m = D+1$ can be replaced by (A.3'). This is proved in a manner analogous to that for Examples 1 and 2 using the fact that $\sum_{h \in I_{++}^D} (h'1)^{-m} < \infty$ if $m \geq D+1$, where I_{++}^D denotes the set of positive integers.

If the true model in Example 1 is a linear regression model with regressors $(x_{ij} : j = 1, \dots, h^*)$ for some h^* finite and $p_n \geq h^*$ for all n large, then $\inf_{h \in H_n} R_n(h) = O(1/n)$ and Assumption (A.3') does not hold. Thus,

(A.3') holds in Example 1 only if the linear models that are under consider-

ation in the model selection problem are all just approximations to the true model. Similarly, in Examples 2 and 3, (A.3') holds only if $f(\cdot)$ does not have a finite expansion in terms of the series functions $\{x_j(\cdot)\}$ or $\{x_{abc}(\cdot)\}$ (since $\inf_{h \in H_n} R_n(h) - R_n(h^*) = O(1/n)$ for some $h^* < \infty$ if it does and if $\lim_{n \rightarrow \infty} \max\{h : h \in H_n\} \geq h^*$).

COROLLARY 2.1*. In Examples 1 and 2, GC_L is asymptotically optimal if (A.2*) with $m = 2$ and (A.3') hold. In Example 3, GC_L is asymptotically optimal if (A.2*) with $m = D+1$ and (A.3') hold.

EXAMPLE 4 (cont.). As shown by Li (1985, Lemma 4.1), in this example condition (A.1) is implied by the following assumptions on the weights:

$$\begin{aligned} &\text{There exists a positive number } \delta' \text{ such that } w_{n,h}(1) \leq 1 - \delta' \\ &\text{for all } n, h \geq 2. \end{aligned} \quad (2.7)$$

$$\text{For all } n, h, \text{ and } i, w_{n,h}(i) \geq w_{n,h}(i+1) \geq 0. \quad (2.8)$$

$$\sum_{i=1}^h w_{n,h}(i) = 1. \quad (2.9)$$

In addition, (A.3) is implied by

$$\lim_{n \rightarrow \infty} (\inf_{h \in H_n} R_n(h)) n^{1-1/m} = \infty. \quad (A.3'')$$

Thus, we obtain

COROLLARY 2.2*. In Example 4, GC_L is asymptotically optimal if (A.2*), (A.3''), and (2.7)-(2.9) hold.

In this example, the second summand of the GC_L criterion simplifies to $2w_{n,h}(1)n^{-1}\sum_{i=1}^n\sigma_i^2$. If Ω is unknown, then one can replace $n^{-1}\sum_{i=1}^n\sigma_i^2$ ($=\tau_n$) by a consistent estimator, call it $\hat{\tau}_n$. If the weights satisfy

$$\overline{\lim}_{n \rightarrow \infty} \sup_{h \in H_n} \frac{w_{n,h}(1)}{\sum_{i=1}^h w_{n,h}^2(i)} < \infty, \quad (2.10^*)$$

then GC_L based on $\hat{\tau}_n$ still is asymptotically optimal. This follows because

$$\sup_{h \in H_n} \frac{|\hat{\tau}_n - \tau_n| w_{n,h}(1)}{R_n(h)} \leq \sup_{h \in H_n} \frac{|\hat{\tau}_n - \tau_n| w_{n,h}(1)}{\inf_{j \geq 1} \sigma_j^2 \sum_{i=1}^h w_{n,h}^2(i)}.$$

For example, if the weights satisfy (2.7)-(2.9) and

$$w_{n,h}(1) \leq Ch^{-1} \text{ for some } C < \infty, \text{ for all } h \in H_n, n \geq 1, \quad (2.11^*)$$

and H_n is a subset of $\{2, \dots, n\}$, then (2.10*) holds (since $\sum_{i=1}^h w_{n,h}^2(i) \geq w_{n,h}^2(1) + h^{-1}(1 - w_{n,h}(1))^2$). The former conditions are easily seen to hold for common weights, such as uniform, triangular, and quadratic weights.

COROLLARY 2.3*. In Example 4, if $\tau_n = n^{-1} \sum_{i=1}^n \sigma_i^2$ is replaced in GC_L by an estimator $\hat{\tau}_n$ such that $\hat{\tau}_n - \tau_n \xrightarrow{P} 0$, then GC_L is still asymptotically optimal under (A.2*), (A.3"), (2.7)-(2.9), and (2.10*). (2.10*) can be replaced by (2.11*) if H_n is a subset of $\{2, \dots, n\}$ for all n .

REMARK. An analogous result holds in the kernel nonparametric regression estimation example when the smoothing parameter is chosen from a finite but expanding grid of points.

3. Generalized cross-validation

Li (1987) introduced the nil-trace estimator as a tool for establishing the asymptotic optimality of GCV based on the asymptotic optimality of C_L . Here we use the same tool to obtain conditions for the asymptotic optimality of GCV based on GC_L when the model errors are heteroskedastic. These conditions are more restrictive than in the homoskedastic error case and do not cover all of the examples.

The following conditions are used by Li (1987) in the homoskedastic error case and will be used here:

$$\inf_{h \in H_n} L_n(h) \stackrel{P}{\rightarrow} 0; \quad (\text{A.4})$$

$$\begin{aligned} &\text{For any sequence } (h_n \in H_n) \text{ such that } n^{-1} \text{tr } M_n(h)M_n'(h) \rightarrow 0, \\ &\text{we have } (n^{-1} \text{tr } M_n(h_n))^2 / n^{-1} \text{tr } M_n(h_n)M_n'(h_n) \rightarrow 0; \end{aligned} \quad (\text{A.5})$$

$$\sup_{h \in H_n} |n^{-1} \text{tr } M_n(h)| \leq \gamma_1 \text{ for some } 0 < \gamma_1 < 1; \text{ and} \quad (\text{A.6})$$

$$\sup_{h \in H_n} (n^{-1} \text{tr } M_n(h))^2 / n^{-1} \text{tr } M_n(h)M_n'(h) \leq \gamma_2 \text{ for some } 0 < \gamma_2 < 1. \quad (\text{A.7})$$

(In Li (1987), (A.6) is stated without the absolute value signs. Inspection of his proof shows that they should be added. This change has little impact on the restrictiveness of the assumption.)

Assumption (A.4) requires the existence of a consistent choice of $\{h_n : n \geq 1\}$ when μ_n is known. This is not overly restrictive. By Markov's inequality, (A.4) is satisfied if $\inf_{h \in H_n} R_n(h) \rightarrow 0$ as $n \rightarrow \infty$. In Example 1,

(A.4) requires that the true model can be approximated arbitrarily well by a

linear model with a sufficiently large number of regressors. In Examples 2-4, (A.4) requires a weak form of consistency of the nonparametric estimator under consideration for some sequence $\{h_n : n \geq 1\}$. Andrews and Whang (1989) provide conditions under which (A.4) holds in Examples 2 and 3. See below for comments on (A.5)-(A.7).

For the heteroskedastic error case, we need two additional conditions:

$$\sup_{h \in H_n} \left[\left| n^{-1} \text{tr } M_n(h) \Omega - n^{-1} \text{tr } M_n(h) n^{-1} \text{tr } \Omega \right| / \left[(1 - n^{-1} \text{tr } M_n(h)) R_n(h) \right] \right] \rightarrow 0; \quad (\text{H.1}^*)$$

$$m_i(h) \geq 0 \quad \forall i = 1, \dots, n, \quad \forall h \in H_n. \quad (\text{H.2}^*)$$

(As above, $m_i(h)$ is the i -th diagonal element of $M_n(h)$.) Assumption (H.2*) is not very restrictive; it is satisfied in Examples 1-4. Assumption (H.1*), however, is restrictive. It is satisfied if the diagonal elements of $M_n(h)$ are all equal, since $n^{-1} \text{tr } M_n(h) \Omega - n^{-1} \text{tr } M_n(h) n^{-1} \text{tr } \Omega$ in this case. Thus, in Example 4, (H.1*) is satisfied, but in Examples 1-3 it is not necessarily satisfied. If (H.1*) does not hold, then the GCV criterion differs from $L_n(h)$ by a term that depends on h and is not negligible asymptotically relative to $L_n(h)$.

THEOREM 3.1*. *Under assumptions (A.1), (A.2*), (A.3)-(A.7), (H.1*), and (H.2*), \hat{h}_G is asymptotically optimal.*

EXAMPLE 4 (cont.). Consider nearest neighbor weights $w_{n,h}(\cdot)$ that satisfy (2.7)-(2.9) and

$$\begin{aligned} &\text{There exist fixed positive numbers } \lambda_1 \text{ and } \lambda_2 \text{ such that} \\ &w_{n,h}(1) \leq \lambda h^{-(1/2+\lambda_2)} \quad \text{for all } h \in H_n, \quad n \geq 1. \end{aligned} \quad (3.9)$$

Condition (3.9) is satisfied by most commonly used weights. As shown in Li (1987, p. 967), (2.7) and (3.9) imply both (A.5) and (A.7). Since GCV is undefined when $h = 1$, we take H_n to be some subset of $\{2, \dots, n\}$. In this case, (A.6) reduces to (2.7). In addition, (H.2*) follows from (2.8) and (H.1*) holds by the definition of $M_n(h)$. Hence, we get the following corollary to Theorem 3.1*:

COROLLARY 3.1*. *In Example 4, suppose the nearest neighbor weights satisfy (2.7)-(2.9) and (3.9). Then, \hat{h}_G is asymptotically optimal if (A.2*), (A.3"), and (A.4) hold and H_n is some subset of $\{2, \dots, n\}$.*

REMARKS. 1. The conditions of the Corollary are exactly the same as those of Li's (1987) Corollary 3.2 except the errors are allowed to be heteroskedastic. A similar generalization of Li's results for Examples 1-3 does not hold, because (H.1*) is not generally satisfied in the latter examples.

2. In the treatment of problems with continuous index sets H_n , one also needs a condition such as (H.1*) in order to establish the asymptotic optimality of GCV. Note that for kernel estimators of nonparametric regression models, (H.1*) is satisfied. This is consistent with Härdle, Hall, and Marron's (1988) results for GCV using kernel estimators. Also note that (H.1*) is not satisfied by local regression, spline, or ridge regression estimators. It would be useful to quantify the extent of the potential asymptotic non-optimality of GCV for such estimators.

4. Cross-validation

Let $\underline{\mu}_n^c(h) = \tilde{M}_n(h)y_n$ denote the delete-one estimator of $\underline{\mu}_n$. By construction, $\tilde{M}_n(h)$ has all diagonal elements equal to zero. Hence, the CV choice of h is just the GC_L choice of h based on the delete-one estimator $\underline{\mu}_n^c(h)$. Using this observation, the asymptotic optimality of GC_L can be used to obtain the asymptotic optimality of CV, as in Li (1987).

Let

$$\tilde{L}_n(h) = n^{-1} \|\underline{\mu}_n - \underline{\mu}_n^c(h)\|^2 \quad \text{and} \quad \tilde{R}_n(h) = E\tilde{L}_n(h) .$$

THEOREM 4.1*. Suppose (A.1), (A.2*), (A.3), (A.4), and the following conditions hold:

$$\overline{\lim}_{n \rightarrow \infty} \sup_{h \in H_n} \lambda(\tilde{M}_n(h)) < \infty , \quad (\text{A.8})$$

$$\sum_{h \in H_n} (n\tilde{R}_n(h))^{-m} \rightarrow 0, \quad \text{and} \quad (\text{A.9})$$

$$\begin{aligned} \text{For any sequence } \{h_n \in H_n\}, \text{ we have } \tilde{R}_n(h_n)/R_n(h_n) \rightarrow 1 \\ \text{if either } R_n(h_n) \rightarrow 0 \text{ or } \tilde{R}_n(h_n) \rightarrow 0. \end{aligned} \quad (\text{A.10})$$

Then \hat{h}_c is asymptotically optimal.

REMARK. This Theorem is a direct analogue of Li's (1987) Theorem 5.1.

EXAMPLES 1-3 (cont.). In these examples, (A.8)-(A.10) are implied by (A.2*) (with $m = 2$ in Examples 1 and 2 and $m = D+1$ in Example 3) and (A.3') plus

$$\overline{\lim}_{n \rightarrow \infty} \sup_{h \in H_n} \bar{\lambda}(M_n(h)) < 1 \text{ and} \quad (5.1)$$

There exists a positive constant $\Lambda < \infty$ such that

$$\text{for all } h \in H_n \text{ and } n \geq 1, \bar{\lambda}(M_n(h)) \leq \Lambda n^{-1} \text{tr } M_n(h), \quad (5.2)$$

where $\bar{\lambda}(\cdot)$ denotes the largest diagonal element of a matrix.

Condition (5.1) requires the self-weights $\{m_i(h)\}$ to be bounded away from one. (They are necessarily ≤ 1 , since $M_n(h)$ is a projection matrix.) This condition is not overly restrictive, since its failure indicates potentially extreme overfitting of the model. If some self-weight $m_i(h)$ is close to one, then the non-diagonal elements of the i -th row of $M_n(h)$ must be close to zero, the estimator $\hat{\mu}_i(h)$ of μ_i must be close to y_i , and the delete-one estimator of μ_i may deviate substantially from $\hat{\mu}_i(h)$. In this scenario, the CV criterion cannot be expected to perform well.

Condition (5.2) prohibits highly unbalanced designs. It is equivalent to requiring the ratio of the maximum to the average diagonal element of $M_n(h)$ to be bounded above by some $\Lambda < \infty$ for all $h \in H_n$ and $n \geq 1$. If (5.2) does not hold, then some elements of μ_n are estimated much less accurately than others using $\hat{\mu}_n(h)$, since the variance of $\hat{\mu}_i(h)$ equals $\sigma_i^2 m_i(h)$.

THEOREM 4.2*. In Examples 1 and 2, if (A.2*) with $m = 2$, (A.3'), (A.4), (5.1), and (5.2) hold, then \hat{h}_c is asymptotically optimal. In Example 3, the same conditions but with $m = D+1$ suffice for asymptotic optimality of \hat{h}_c .

REMARK. The Theorem shows that CV is asymptotically optimal in Examples 1-3 under the same conditions when the errors are heteroskedastic as when they are homoskedastic. This contrasts with the results of Section 3 for GCV. For these examples, the asymptotic optimality of GCV does not carry over from homoskedastic to heteroskedastic errors.

EXAMPLE 4 (cont.). Consider the following assumption on the regression function:

$$f_{\infty} = \sup_{z \in Z} |f(z)| < \infty. \quad (\text{F.1}^*)$$

Using this assumption and Theorem 4.1*, we get the following result for the use of CV with nearest neighbor estimators:

THEOREM 4.3*. In Example 4, if the nearest neighbor weights satisfy (2.7)-(2.9) and (3.9). Then, \hat{h}_c is asymptotically optimal if (A.2*), (A.3"), (A.4), and (F.1*) hold and H_n is some subset of $\{2, \dots, n\}$.

REMARK. In this example as well, CV is asymptotically optimal with heteroskedastic errors under the same conditions as with homoskedastic errors.

5. Proofs

PROOF OF THEOREM 2.1*. The proof of (1.6) is the same as Li's (1987) proof of Theorem 2.1 except for the following: $\sigma^2 \text{tr } M_n(h)$ is replaced by $\text{tr } M_n(h)\Omega$ in (2.1)-(2.3) and everywhere it appears in the proof of Theorem 2.1, $\sigma^2 n^{-1} \text{tr } M_n(h)M_n'(h) \leq R_n(h)$ is replaced by $(\inf_{i \geq 1} \sigma_i^2 / \sup_{i \geq 1} \sigma_i^2) \times n^{-1} \text{tr } M_n(h)M_n'(h) \leq R_n(h)$ just above (6.1), and $\sigma^2 \text{tr } M_n(h)M_n'(h)$ is replaced by $\text{tr } M_n'(h)M_n(h)\Omega$ in (6.2). Li's proof uses Theorem 2 of Whittle

(1960). The latter also applies when the errors are heteroskedastic provided (A.2*) holds.

The second result of the Theorem, (1.7*), holds by (1.6) above and

$$\sup_{h \in H_n} |L_n(h)/R_n(h) - 1| \stackrel{P}{\rightarrow} 0. \quad (5.1^*)$$

Equation (5.1*) is the same as (2.4) of Li (1987). Its proof in the context of heteroskedastic errors is included in the proof of the preceding paragraph, because (2.4) is established as part of Li's proof of Theorem 2.1.

Note that the summand $2n^{-1} \text{tr } M_n(h)\Omega$ arises in the GC_L criterion because it equals $2n^{-1} E e_n' M_n(h) e_n$, where $e_n = (e_1, \dots, e_n)'$. \square

In view of (5.1*), whenever (1.6) holds for some estimator \hat{h} , so does (1.7*), provided (A.1), (A.2*), and (A.3) are in force. Since the latter (or assumptions that imply the latter) are assumed in each of the results of this paper, it suffices to establish just (1.6) in the remainder of this section.

Following Li (1987, p. 965) define the nil-trace estimator $\bar{\mu}_n(h)$ as

$$\bar{\mu}_n(h) = -\alpha y_n + (1+\alpha)\hat{\mu}_n(h), \quad (5.2^*)$$

where $\alpha = n^{-1} \text{tr } M_n(h) / (1 - n^{-1} \text{tr } M_n(h))$. The matrix $\bar{M}_n(h)$ associated with $\bar{\mu}_n(h)$ is given by

$$\bar{M}_n(h) = -\alpha I_n + (1+\alpha)M_n(h). \quad (5.3^*)$$

It has trace equal to zero. Define

$$\bar{L}_n(h) = n^{-1} \|\mu_n - \bar{\mu}_n(h)\|^2 \quad \text{and} \quad \bar{R}_n(h) = E\bar{L}_n(h). \quad (5.4^*)$$

PROOF OF THEOREM 3.1*. Let \bar{h}_G denote the GC_L choice of h based on $\bar{\mu}_n(h)$. That is, \bar{h}_G minimizes

$$n^{-1} \|y_n - \bar{\mu}_n(h)\|^2 + n^{-1} \text{tr } \bar{M}_n(h)\Omega \quad (5.5^*)$$

over H_n , where

$$n^{-1} \text{tr } \bar{M}_n(h)\Omega = (n^{-1} \text{tr } M_n(h)\Omega - n^{-1} \text{tr } M_n(h)n^{-1} \text{tr } \Omega) / (1 - n^{-1} \text{tr } M_n(h)) . \quad (5.6^*)$$

Using Li's (1987) proof of Theorem 3.2, we find that \bar{h}_G is asymptotically optimal for use with the estimator $\hat{\mu}_n(h)$. The following changes are needed in Li's proof: All references to \hat{h}_G are changed to \bar{h}_G . The expression for $n\bar{R}_n(h)$ is replaced by

$$\begin{aligned} n\bar{R}_n(h) = & [\|A_n(h)\bar{\mu}_n\|^2 + \text{tr } M'_n(h)M_n(h)\Omega - 2n^{-1}(\text{tr } M_n(h))\text{tr } M_n(h)\Omega \\ & + n^{-2}(\text{tr } M_n(h))^2 \text{tr } \Omega] / (n^{-1} \text{tr } A_n(h))^2 . \end{aligned} \quad (5.7^*)$$

Then, Li's (6.3) follows from (A.6) and (A.7), and his (6.4) follows from (A.6), (A.7), (H.2*), and (A.2*) (where $R_n(h)$ in (6.4) is as defined in the present paper). Assumption (H.2*) is used here to bound the magnitude of $|\text{tr } M_n(h)\Omega|$. The rest of Li's proof of Theorem 3.2 follows without change. Li's use of his Theorem 3.1 is justified in the present context, because it holds as stated provided (A.2*) is assumed. His proof of Theorem 3.1 goes through with heteroskedastic errors, since $R_n(h) \geq \inf_{i \geq 1} \sigma_i^2 n^{-1} \text{tr } M'_n(h)M_n(h)$. This completes the proof of the asymptotic optimality of \bar{h}_G . Note that (H.1*) has not been used thus far.

We now show that

$$L_n(\hat{h}_G)/L_n(\bar{h}_G) \stackrel{P}{\rightarrow} 1. \quad (5.8^*)$$

This result plus the asymptotic optimality of \bar{h}_G gives the desired result.

To show (5.8*), note that

$$n^{-1} \|y_n - \hat{\mu}_n(h)\|^2 / (1 - n^{-1} \text{tr } M_n(h))^2 = n^{-1} \|y_n - \bar{\mu}_n(h)\|^2. \quad (5.9^*)$$

Hence, \bar{h}_G also can be defined as the value that minimizes

$$n^{-1} \|y_n - \hat{\mu}_n(h)\|^2 / (1 - n^{-1} \text{tr } M_n(h))^2 + n^{-1} \text{tr } \bar{M}_n(h) \Omega. \quad (5.10^*)$$

In analogy with Li's (1987) argument of (2.1)-(2.4) and in view of (5.6*), (H.1*), and (5.1*), we obtain (5.8*). \square

PROOF OF THEOREM 4.1*. The proof is the same as Li's (1987) proof of Theorem 5.1 except the appeals to Theorems 2.1 and 3.2 are replaced by appeals to Theorems 2.1* and 3.1*. \square

PROOF OF THEOREM 4.2*. It suffices to establish (A.8)-(A.10), then Theorem 4.1* yields the desired result. First consider Examples 1 and 2. Li's (1987) proof of Theorem 5.2 shows that (A.1) and (5.1) imply (A.8). Li's proof that (A.9) and (A.10) hold also applies in the present case provided one replaces $\sigma^2 h_n^{-1}$, $\sigma^2 \text{tr } \bar{M}_n(h_n) \bar{M}'_n(h_n)$, $\sigma^2 (1 + o(1)) \text{tr } M_n(h) M'_n(h)$, and $\sigma^2 \bar{h}_n$ in his proof by $\inf_{i \geq 1} \sigma_i^2 h_n^{-1}$, $\text{tr } \bar{M}'_n(h_n) \bar{M}_n(h_n) \Omega$, $(1 + o(1)) \text{tr } M'_n(h) M_n(h) \Omega$, and $\inf_{i \geq 1} \sigma_i^2 \bar{h}_n$, respectively, and provided one replaces σ^{-4} in his equation (2.6) by $1/\inf_{i \geq 1} \sigma_i^4$.

The proof for Example 3 is similar, even though h is vector-valued. In those places where h_n or \bar{h}_n is used as a scalar in Li's proof of Theorem

5.2, it needs to be replaced by $\text{tr } M_n(h_n)$ or $\text{tr } M_n(\bar{h}_n)$ respectively. The condition on m in Example 3 arises because the analogue of (2.6) of Li (1987) needed for this example in Li's proof of Theorem 5.2 holds only if

$$\sum_{h \in I_{++}^D} (h'1)^{-m} < \infty \quad (\text{as in Section 2 above}) \quad \text{and the latter holds when } m = D+1. \square$$

PROOF OF THEOREM 4.3*. It suffices to show that (A.8)-(A.10) hold. (A.8) and (A.9) hold by the same argument as given by Li (1987, proof of Theorem 5.3).

We now show that (A.10) holds. By Li's proof of Theorem 5.3, we have

$$|n^{-1} \|\underline{\mu}_n - M_n(h)\underline{\mu}_n\|^2 - n^{-1} \|\underline{\mu}_n - \tilde{M}_n(h)\underline{\mu}_n\|^2| \leq 4w_{n,h}(1)^2 f_\infty^2. \quad (5.11^*)$$

In addition, it is straightforward to see that

$$\begin{aligned} n^{-1} \text{tr } M_n(h) \Omega M_n'(h) &= n^{-1} \sum_{i=1}^n \sum_{m=1}^n \sigma_v^2 \sum_{j=1}^h w_{n,h}(j)^2 1_{(m=i(j))} \quad \text{and} \\ n^{-1} \text{tr } \tilde{M}_n(h) \Omega \tilde{M}_n'(h) &= n^{-1} \sum_{i=1}^n \sum_{m=1}^n \sigma_v^2 \sum_{j=1}^{h+1} w_{n,h}(j-1)^2 1_{(m=i(j))}, \end{aligned} \quad (5.12^*)$$

where $w_{n,h}(0) = 0$. Thus, we get

$$n^{-1} \text{tr } M_n(h) \Omega M_n'(h) \geq \inf_{v \geq 1} \sigma_v^2 \sum_{j=1}^h w_{n,h}(j)^2, \quad \text{and} \quad (5.13^*)$$

$$\begin{aligned} &|n^{-1} \text{tr } M_n(h) \Omega M_n'(h) - n^{-1} \text{tr } \tilde{M}_n(h) \Omega \tilde{M}_n'(h)| \\ &\leq \sup_{v \geq 1} \sigma_v^2 n^{-1} \sum_{i=1}^n \sum_{m=1}^n \sum_{j=1}^{h+1} 1_{(m=i(j))} |w_{n,h}(j-1)^2 - w_{n,h}(j)^2| \quad (5.14^*) \\ &= \sup_{v \geq 1} \sigma_v^2 2w_{n,h}(1)^2, \end{aligned}$$

where the equality uses (2.8).

Since $R_n(h) = n^{-1} \|\underline{\mu}_n - M_n(h)\underline{\mu}_n\|^2 + n^{-1} \text{tr } M_n(h)\Omega M_n'(h)$ and analogously for $\tilde{R}_n(h)$, we get

$$|R_n(h) - \tilde{R}_n(h)| \leq (4f_\infty^2 + 2 \sup_{i \geq 1} \sigma_i^2) w_{n,h}(1)^2 . \quad (5.15^*)$$

Thus, the desired result $|R_n(h) - \tilde{R}_n(h)|/R_n(h) \stackrel{P}{\rightarrow} 0$ holds if

$$w_{n,h}(1)^2 / \sum_{j=1}^h w_{n,h}(j)^2 \stackrel{P}{\rightarrow} 0 . \quad (5.16^*)$$

Under (2.7) and (3.9), this holds by the same argument as given by Li (1987, proof of Theorem 5.3). (Note that Li's proof of this contains two typographical errors--on the third last line of p. 974, h^{-1} should be $(\delta')^2 h^{-1}$ and (3.8) should be (3.9).) \square

FOOTNOTE

*The author gratefully acknowledges the support of the National Science Foundation and the Alfred P. Sloan Foundation through grant No. SES-8618617 and a Research Fellowship respectively.

REFERENCES

- Allen, D. M., 1974, The relationship between variable selection and data augmentation and a method for prediction, *Technometrics* 16 125-127.
- Andrews, D. W. K., 1988, Asymptotic normality of series estimators for various nonparametric and semiparametric models, Discussion Paper No. 874 (Cowles Foundation, Yale University).
- Andrews, D. W. K. and Whang, Y. J., 1989, Additive interactive regression models: Lifting the curse of dimensionality, unpublished manuscript (Cowles Foundation, Yale University).
- Bierens, H. J., 1987, Kernel estimators of regression functions, in: T. F. Bewley, ed., *Advances in Econometrics: Fifth World Congress, Vol. I* (Cambridge University Press, New York).
- 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.
- Gallant, A. R., 1981, On the bias in flexible functional forms and an essentially unbiased form, *Journal of Econometrics* 15 211-245.
- Geisser, S., 1975, The predictive sample reuse method with applications, *Journal of the American Statistical Association* 70 320-328.
- Geman, S. and Hwang, C. R., 1982, Nonparametric maximum likelihood estimation by the method of sieves, *Annals of Statistics* 10 401-414.
- Härdle, W., Hall, P., and Marron, J. S., 1988, How far are automatically chosen regression smoothing parameters from their optimum? (with discussion), *Journal of the American Statistical Association* 83 86-101.

- Li, K.-C., 1985, From Stein's unbiased risk estimates to the method of generalized cross-validation, *Annals of Statistics* 13 1352-1377.
- Li, K.-C., 1987, Asymptotic optimality for C_p , C_L , cross-validation, and generalized cross-validation: Discrete index set, *Annals of Statistics* 15 958-975.
- Mallows, C. L., 1973, Some comments on C_p , *Technometrics* 15 661-675.
- Stone, M., 1974, Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society, Series B* 36 111-147.
- Wahba, G., 1986, Partial and interaction splines for the semiparametric estimation of functions of several variables, in: T. E. Boardman, ed., *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface* (American Statistical Association, Washington) 75-80.
- Wahba, G., 1989, *Spline Models in Statistics*, Regional conference series in applied mathematics (Society for Industrial and Applied Mathematics, Philadelphia) forthcoming.
- Wahba, G. and Wold, S., 1975, A completely automatic French curve: Fitting spline functions by cross-validation, *Communications in Statistics* 4 1-17.
- Whittle, P., 1960, Bounds for the moments of linear and quadratic forms in independent variables, *Theory of Probability and its Applications* 5 302-305.