Cowles Foundation Paper 790

Journal of Econometrics 47 (1991) 359–377. North-Holland

Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors

Donald W. K. Andrews*

Yale University, New Haven, CT 06520, USA

Received May 1989, final version received January 1990

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 A feasible version of GC_L, however, is available only in some applications. 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 Thus, the CV criterion is the only feasible criterion of the three that is asymptotically optimal under general conditions. The proofs rely heavily on results of Li (1987).

1. Introduction

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

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

*The author thanks two referees for helpful comments and suggestions In addition, the author gratefully acknowledges the support of the National Science Foundation and the Alfred P Sloan Foundation through Grant Nos. SES-8618617 and SES-8821021 and a Research Fellowship, respectively

where $\boldsymbol{\mu}_n = (\mu_1, \ldots, \mu_n)'$ is the unobserved mean vector of $\boldsymbol{y}_n$ and $\boldsymbol{e}_n = (e_1, \ldots, e_n)'$ is the unobserved error vector comprised of independent, mean zero, variance σ_i^2 errors. Consider a class of linear estimators $\hat{\boldsymbol{\mu}}_n(h) = M_n(h)\boldsymbol{y}_n$, where each estimator is indexed by a parameter h in an index set $\mathscr{H}_n$. 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 $\boldsymbol{y}_n$ to select $\hat{h}$ from $\mathscr{H}_n$ in such a way as to make the average squared error,

$$L_n(\hat{h}) = n^{-1}\|\boldsymbol{\mu}_n - M_n(\hat{h})\,\boldsymbol{y}_n\|^2,$$

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 $x_{i1}, \ldots, x_{ip_n}$ 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 $\boldsymbol{\mu}_n$ using the least squares estimator $\hat{\boldsymbol{\mu}}_n(h) = X_h(X_h'X_h)^{-1}X_h'\boldsymbol{y}_n$. 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 $\mathscr{H}_n$ is $\{1, \ldots, p_n\}$. The goal is to determine an appropriate model for the purpose of estimating $\boldsymbol{\mu}_n$.

Example 2. Series Estimation of a Nonparametric Regression Model [Gallant (1981), Geman and Hwang (1982), Andrews (1991)]: The mean of y_i is an unknown function $f(\cdot)$ of an observed regressor vector z_i in $\mathscr{Z} \subset R^p$, $\mu_i = f(z_i)$. 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 $x_j(\cdot)$, $j = 1, \ldots, h$, are known (e.g., trigonometric functions) and the coefficients β_j, $j = 1, \ldots, h$, are unknown. One estimates $\boldsymbol{\mu}_n$ by estimating the unknown constants $\{\beta_j\}$ by least squares: $\hat{\boldsymbol{\mu}}_n(h) = X_h(X_h'X_h)^{-1}X_h'\boldsymbol{y}_n$, 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 $\mathscr{H}_n$ is $\{1, \ldots, n\}$ or some subset thereof. The goal is to determine the appropriate number of terms in the series expansion to be used in estimating $\boldsymbol{\mu}_n$.

Example 3. Series Estimation of an Additive Interactive Regression (AIR) Model [Andrews (1991), Andrews and Whang (1990)]: 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 $f_{ab}(z_t)$ depends on only 'a' ($\leq d$) different elements of the d-vector z_t for each $b = 1, \dots, B(a)$. For example, one might have $f_{1b}(z_t) = f_{1b}^*(z_{tb})$ and $f_{2b}(z_t) = f_{2b}^*(z_{t1}, z_{t2})$, where $z_t = (z_{t1}, \dots, z_{td})'$. If $A = 1$, the model is an additive regression model. If $A > 1$, the model allows interactions between the elements of z_t. 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 (1990), 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 $f(z_t)$ is constructed using a series approximation $\sum_{c=1}^{h_{ab}} x_{abc}(z_t)\beta_{abc}$ of each function $f_{ab}(z_t)$, 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_t as does $f_{ab}(\cdot)$ for all $c = 1, \dots, h_{ab}$. One estimates $\boldsymbol{\mu}_n$ by using least squares to estimate $\{\beta_{abc}\}$: $\hat{\boldsymbol{\mu}}_n(h) = X_h(X_h'X_h)^{-1}X_h \boldsymbol{y}_n$, where X_h is the (full rank) $n \times (\sum_{a=1}^{A}\sum_{b=1}^{B(a)} h_{ab})$ matrix with ith row given by the elements of $\{x_{abc}(z_t)$: $c = 1, \dots, h_{ab}$, $b = 1, \dots, B(a)$, $a = 1, \dots, A\}$. The parameter h in this example is a vector $(h_{11}, \dots, h_{1B(1)}, h_{21}, \dots, h_{AB(A)})'$ of nonnegative integers of dimension $D = \sum_{a=1}^{A} B(a)$. The index set $\mathscr{H}_n$ is some subset of $\{h \in I_+^D$: $h'\mathbf{1} \leq n\}$, where I_+ denotes the set of nonnegative integers and $\mathbf{1}$ 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 $z_{t(j)}$ denote the jth nearest neighbor of z_t in the sense that $\|z_t - z_{t(j)}\|$ is the jth smallest number among the n values $\|z_t - z_v\|$, $v = 1, \dots, n$. (Ties may be broken in any systematic fashion.) For a given weight function $w_{n,h}(\cdot)$, the h-nearest-neighbor estimate of μ_t is $\hat{\mu}_t(h) = \sum_{j=1}^{h} w_{n,h}(j) y_{t(j)}$. Hence, $\hat{\boldsymbol{\mu}}_n(h)$ is of the form $M_n(h)\boldsymbol{y}_n$, 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 $\mathscr{H}_n$ is $\{1, \dots, n\}$ or some subset thereof. The goal is to use $\boldsymbol{y}_n$ to determine the number of neighbors to include in the estimate of $\boldsymbol{\mu}_n$.

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 (1990)], 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

$$\min_{h \in \mathscr{H}_n} \left(n^{-1} \| y_n - \hat{\mu}_n(h) \|^2 + 2n^{-1} \operatorname{tr} M_n(h) \Omega \right), \tag{1.1*}$$

where Ω is an $n \times n$ diagonal matrix with diagonal elements $\{\sigma_1^2, \ldots, \sigma_n^2\}$. 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

$$\min_{h \in \mathscr{H}_n} \frac{n^{-1} \| y_n - \hat{\mu}_n(h) \|^2}{\left(1 - n^{-1} \operatorname{tr} M_n(h)\right)^2}, \tag{1.2}$$

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

(iii) (Delete-one) Cross-validation (CV) [Schmidt (1971), 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 $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$ 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

$$\min_{h \in \mathscr{H}_n} n^{-1} \| y_n - \tilde{M}_n(h) y_n \|^2, \tag{1.3}$$

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 $\{\sigma_1^2, \ldots, \sigma_n^2\}$ 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,

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

where $D_n(h)$ is an $n \times n$ diagonal matrix with ith diagonal element equal to $(1 - m_i(h))^{-1}$, $m_i(h)$ is the ith 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

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

In Examples 1–4, $n^{-1} \operatorname{tr} M_n(h)$ equals h/n, h/n, $h'\mathbf{1}/n$, and $w_{n,h}(1)$ respectively. In Example 4,

$$n^{-1} \operatorname{tr} M_n(h)\Omega = w_{n,h}(1) n^{-1} \sum_{n}^{i=1} \sigma_i^2, \quad \hat{y}_{-i} = \sum_{h}^{j=1} w_{n,h}(j) y_{i(j+1)},$$

and $\tilde{M}_n(h)$ has rows that are permutations of $(w_{n,h}(1), \ldots, w_{n,h}(h), 0, \ldots, 0)$ and diagonal elements that are zeroes.

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

$$\frac{L_n(\hat{h})}{\inf_{h \in \mathscr{H}_n} L_n(h)} \xrightarrow{p} 1 \tag{1.6}$$

and

$$\frac{R_n(\hat{h})}{\inf_{h \in \mathscr{H}_n} R_n(h)} \xrightarrow{p} 1 \tag{1.7*}$$

even when the errors are heteroskedastic, where $R_n(h) = \mathrm{E} L_n(h)$ and '$\xrightarrow{p} 1$' denotes convergence in probability as $n \to \infty$.

In section 2 we find that analogues of Li's (1987) conditions for the asymptotic optimality of Mallows' C_L procedure when the errors are homoskedastic 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. In section 5, we summarize the results of the paper as applied to Examples 1–4. Section 6 contains several conjectures regarding the possible extension of the results of the paper. Section 7 contains proofs of the results of the paper.

2. Generalized C_L

We have

$$R_n(h) = \mathrm{E}L_n(h) = n^{-1}\|A_n(h)\boldsymbol{\mu}_n\|^2 + n^{-1}\operatorname{tr} M_n'(h)M_n(h)\Omega,$$

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

$$\overline{\lim_{n\to\infty}} \sup_{h\in\mathscr{H}_n} \lambda(M_n(h)) < \infty, \tag{A.1}$$

$$\sup_{i\geq 1} \mathrm{E}e_i^{4m} < \infty, \qquad 0 < \inf_{i\geq 1} \sigma_i^2 \leq \sup_{i\geq 1} \sigma_i^2 < \infty, \tag{A.2*}$$

$$\sum_{h\in\mathscr{H}_n} (nR_n(h))^{-m} \to 0 \quad \text{as} \quad n\to\infty, \tag{A.3}$$

for some positive integer m. [Assumption (A.1) of Li (1987, p. 961) has '$\lim_{n\to\infty}$' in place of '$\overline{\lim}_{n\to\infty}$', 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

$$\inf_{h\in\mathscr{H}_n} nR_n(h) \to \infty, \tag{A.3'}$$

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 $\mathrm{tr}M_n(h)\Omega$ 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 $\sum_{h\in I_{++}^D}(h'\mathbf{1})^{-m} < \infty$ if $m \geq D + 1$, where I_{++} denotes the set of positive integers.

If the true model in Example 1 is a linear regression model with regressors $\{x_{ij}\colon j = 1,\dots,h^*\}$ for some h^* finite and $p_n \geq h^*$ for all n large, then $\inf_{h\in\mathscr{H}_n} R_n(h) = O(1/n)$ and Assumption (A.3′) does not hold. Thus, (A.3′) holds in Example 1 only if the linear models that are under consideration in the model selection problem are all just approximations to the true model. In fact, the results of Geweke and Meese (1981, p. 63) imply that when the true model is a linear regression model with $h^* < \infty$ and the errors are iid normal, then the GC_L criterion (equivalently C_p in this case) is not asymptotically optimal in the sense of (1.7*). 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 $\inf_{h\in\mathscr{H}_n} R_n(h) = R_n(h^*) = O(1/n)$ for some $h^* < \infty$ if it does and if $\underline{\lim}_{n\to\infty} \max\{h\colon h\in\mathscr{H}_n\} \geq h^*$].

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:

$$\text{there exists a positive number } \delta' \text{ such that } \quad w_{n,h}(1) \leq 1 - \delta' \quad \text{for all } n,\quad h \geq 2; \tag{2.7}$$

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

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

In addition, (A.3) is implied by

$$\lim_{n\to\infty}\Big(\inf_{h\in\mathscr{H}_n} R_n(h)\Big)n^{1-1/m} = \infty. \tag{A.3″}$$

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 Example 4, a feasible version of GC_L is available. The second summand of the GC_L criterion simplifies to $2w_{n,h}(1)n^{-1}\sum_{i=1}^{n}\sigma_i^2$ in this example. If Ω is unknown, then one can replace $n^{-1}\sum_{i=1}^{n}\sigma_i^2(\equiv \tau_n)$ by a consistent estimator, call it $\hat{\tau}_n$. If the weights satisfy

$$\overline{\lim_{n\to\infty}} \sup_{h\in\mathscr{H}_n} \frac{w_{n,h}(1)}{\sum_{i=1}^{h} w_{n,h}^2(i)} < \infty, \tag{2.10*}$$

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

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

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

$$w_{n,h}(1) \le Ch^{-1} \quad \text{for some} \quad C<\infty, \quad \text{for all} \quad h\in\mathscr{H}_n, \quad n\ge 1, \tag{2.11*}$$

and $\mathscr{H}_n$ is a subset of $\{2,\ldots,n\}$, then (2.10*) holds [since $\sum_{i=1}^{h}w_{n,h}^2(i) \ge w_{n,h}^2(1) + h^{-1}(1-w_{n,h}(1))^2$]. 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 $\mathscr{H}_n$ is a subset of $\{2,\ldots,n\}$ 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:

$$\inf_{h \in \mathscr{H}_n} L_n(h) \xrightarrow{p} 0; \tag{A.4}$$

for any sequence $\{h_n \in \mathscr{H}_n\}$ such that $n^{-1} \operatorname{tr} M_n(h) M_n'(h) \to 0$,

$$\text{we have } \left(n^{-1} \operatorname{tr} M_n(h_n)\right)^2 / n^{-1} \operatorname{tr} M_n(h_n) M_n'(h_n) \to 0; \tag{A.5}$$

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

$$\sup_{h \in \mathscr{H}_n} \left(n^{-1} \operatorname{tr} M_n(h)\right)^2 / n^{-1} \operatorname{tr} M_n(h) M_n'(h) \leq \gamma_2 \quad \text{for some} \quad 0 < \gamma_2 < 1. \tag{A.7}$$

[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 $\boldsymbol{\mu}_n$ is known. This is not overly restrictive. By Markov's inequality, (A.4) is satisfied if $\inf_{h \in \mathscr{H}_n} R_n(h) \to 0$ as $n \to \infty$. 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 (1990) 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:

$$\sup_{h \in \mathscr{H}_n} \left(|n^{-1} \operatorname{tr} M_n(h) \Omega - n^{-1} \operatorname{tr} M_n(h) n^{-1} \operatorname{tr} \Omega| \right. \\ \left. / \left[\left(1 - n^{-1} \operatorname{tr} M_n(h)\right) R_n(h)\right]\right) \to 0; \tag{H.1*}$$

$$m_i(h) \geq 0, \quad \forall i = 1, \ldots, n, \quad \forall h \in \mathscr{H}_n. \tag{H.2*}$$

[As above, $m_i(h)$ is the ith 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 $n^{-1}\operatorname{tr} M_n(h)\Omega = n^{-1}\operatorname{tr} M_n(h) n^{-1}\operatorname{tr}\Omega$ 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

there exist fixed positive numbers λ_1 and λ_2 such that

$$w_{n,h}(1) \le \lambda_1 h^{-(1/2+\lambda_2)} \quad \text{for all} \quad h \in \mathscr{H}_n, \quad n \ge 1. \tag{3.9}$$

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 $\mathscr{H}_n$ to be some subset of $\{2, \ldots, n\}$. 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 $\mathscr{H}_n$ is some subset of $\{2, \ldots, n\}$.*

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 $\mathscr{H}_n$, 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 nonoptimality of *GCV* for such estimators.

4. Cross-validation

Let $\mu_n^c(h) = \tilde{M}_n(h)y_n$ denote the delete-one estimator of μ_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 $\mu_n^c(h)$. Using this observation, the asymptotic optimality of GC_L can be used to obtain the asymptotic optimality of CV, as in Li (1987).

Let

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

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

$$\overline{\lim_{n\to\infty}} \sup_{h\in\mathscr{H}_n} \lambda(\tilde{M}_n(h)) < \infty; \tag{A.8}$$

$$\sum_{h\in\mathscr{H}_n} \left(n\tilde{R}_n(h)\right)^{-m} \to 0; \tag{A.9}$$

for any sequence $\{h_n \in \mathscr{H}_n\}$, we have $\tilde{R}_n(h_n)/R_n(h_n) \to 1$ if either $R_n(h_n) \to 0$ or $\tilde{R}_n(h_n) \to 0$. (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

$$\overline{\lim_{n\to\infty}} \sup_{h\in\mathscr{H}_n} \bar{\lambda}(M_n(h)) < 1, \tag{5.1}$$

there exists a positive constant $A < \infty$ such that, for all $h \in \mathscr{H}_n$ and $n \geq 1$, $\bar{\lambda}(M_n(h)) \leq An^{-1}\,\mathrm{tr}\,M_n(h)$, (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 nondiagonal elements of the ith row of $M_n(h)$ must be close to zero, the estimator $\hat{\mu}_t(h)$ of μ_t must be close to y_t, and the delete-one estimator of μ_t may deviate substantially from $\hat{\mu}_t(h)$. 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 $h \in \mathscr{H}_n$ and $n \geq 1$. If (5.2) does not hold, then some elements of $\boldsymbol{\mu}_n$ are estimated much less accurately than others using $\hat{\boldsymbol{\mu}}_n(h)$, since the variance of $\hat{\mu}_t(h)$ equals $\sigma_t^2 m_t(h)$.

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 [and satisfy (A.2*)] 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:

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

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, suppose 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 $\mathscr{H}_n$ is some subset of $\{2, \ldots, n\}$.*

Remark. In this example as well, *CV* is asymptotically optimal under the same conditions when the errors are heteroskedastic [and satisfy (A.2*)] as when they are homoskedastic.

5. Summary of results for Examples 1–4

Here we summarize the results of the paper with regard to Examples 1–4 under the assumption of heteroskedastic errors. In Example 1 (linear regression model selection), GC_L and *CV* are asymptotically optimal under the conditions given provided none of the models that may be selected is actually true. No asymptotically optimal feasible version of GC_L is available, however, unless additional restrictions are placed on the form of the heteroskedasticity. Also, GCV is not asymptotically optimal.

In Examples 2 (series estimation of a nonparametric regression model) and 3 (series estimation of an additive interactive regression model), GC_L and CV are asymptotically optimal. Here again, no asymptotically optimal feasible version of GC_L is available without further restrictions on Ω and GCV is not asymptotically optimal.

In Example 4 (nearest-neighbor estimation of a nonparametric regression model), GC_L, a feasible version of GC_L, GCV, and CV are each asymptotically optimal under the conditions given.

6. Extensions

In this section we make several conjectures concerning the extension of the above results.

First, consider the case where the errors are correlated, and hence, Ω is nondiagonal. It should be possible to show that GC_L, as defined above, is asymptotically optimal under a suitable set of regularity conditions. [The current proof of Theorem 2.1 does not go through straightforwardly, however, since Whittle's (1960) inequality assumes independence.] On the other hand, obtaining a feasible version of GC_L may require knowledge of the structure of Ω, such as knowledge of Ω up to a consistently estimable finite dimensional parameter. In addition, the GCV criterion will not be asymptotically optimal in general when Ω is nondiagonal. Even in Example 4, GCV will not be asymptotically optimal. Similarly, one would not expect the CV criterion to be asymptotically optimal when Ω is nondiagonal (since the CV criterion is not equivalent to the GC_L criterion based on the delete-one estimator in this case).

Next, consider the case where Ω is diagonal but the Euclidean loss function $L_n(h)$ is replaced by a weighted loss function:

$$L_n(h,W) = \|\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(h)\|_W^2 = (\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(h))'W(\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(h)), \tag{6.1*}$$

where W is a symmetric positive semi-definite matrix of constants. Note that just because $\Omega \neq \sigma^2 I$ does not mean that the Euclidean loss function is inappropriate. The choice of loss function need not be related to the difficulty in estimating the different elements of $\boldsymbol{\mu}_n$ (i.e., to the relative magnitudes of the diagonal elements of Ω). Nevertheless, one still may want to choose a loss function for which $W \neq I$.

When $W \neq I$, none of the criteria discussed above will be asymptotically optimal (unless the difference between W and I goes to zero in some sense as $n \to \infty$). Altered versions of these criteria, however, can be considered. The GC_L criterion can be replaced by

$$n^{-1}\|y_n - \hat{\mu}_n(h)\|_W^2 + 2n^{-1} \operatorname{tr} WM_n(h)\Omega. \tag{6.2*}$$

It should be possible to extend the optimality results for GC_L to the criterion (6.2*) when the loss function is $L_n(h, W)$. If either Ω or W is unknown (e.g., W could equal Ω^{-1}, and hence, be unknown), then a feasible version of (6.2*) is needed. In many cases, some knowledge of the structure of Ω and W (e.g., knowledge of Ω and W up to a consistently estimable finite dimensional parameter) may be needed to obtain such a feasible version.

The GCV criterion can be replaced by

$$\frac{n^{-1}\|y_n - \hat{\mu}_n(h)\|_W}{\left(1 - n^{-1} \operatorname{tr} M_n(h)\right)^2}, \tag{6.3*}$$

when the loss function of interest is $L_n(h, W)$. It should be possible to show that this criterion yields an asymptotically optimal $\hat{h}$ under suitable regularity conditions when (i) W is diagonal with diagonal elements bounded away from zero and infinity and (ii) the diagonal elements of $M_n(h)$ are equal (as in Example 4), but not otherwise. [When conditions (i) and (ii) hold, a Taylor expansion shows that the criteria (6.2*) and (6.3*) are close when $n^{-1} \operatorname{tr} M_n(h)$ is small.]

The CV criterion can be replaced by

$$n^{-1}\|y_n - \tilde{M}_n(h)\, y_n\|_W^2, \tag{6.4}$$

when the loss function of interest is $L_n(h, W)$. It should be possible to show that this criterion is asymptotically equivalent to that of (6.2*), and hence yields an asymptotically optimal $\hat{h}$ for the loss function $L_n(h, W)$, under suitable regularity conditions (that cover Examples 1–4, for example). The criterion (6.4*) does not require knowledge of Ω, but does require knowledge of W. A feasible version of this criterion could be considered if W is estimable.

7. 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: $\sigma^2 \operatorname{tr} M_n(h)$ is replaced by $\operatorname{tr} M_n(h)\Omega$ in (2.1)–(2.3) and everywhere it appears in the proof of Theorem 2.1, $\sigma^2 n^{-1} \operatorname{tr} M_n(h)M_n'(h) \leq R_n(h)$ is replaced by $(\inf_{i\geq 1} \sigma_i^2/\sup_{i\geq 1} \sigma_i^2) \times n^{-1} \operatorname{tr} M_n(h)M_n'(h) \leq R_n(h)$ just above (6.1), and $\sigma^2 \operatorname{tr} M_n(h)M_n'(h)$ is replaced by $\operatorname{tr} M_n'(h)M_n(h)\Omega$ 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

$$\sup_{h \in \mathscr{H}_n} |L_n(h)/R_n(h) - 1| \xrightarrow{p} 0. \tag{7.1*}$$

Eq. (7.1*) 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 $2n^{-1} \operatorname{tr} M_n(h)\Omega$ arises in the GC_L criterion because it equals $2n^{-1}\mathrm{E}e_n' M_n(h) e_n$, where $e_n = (e_1, \ldots, e_n)'$. □

In view of (7.1*), 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 $\bar{\boldsymbol{\mu}}_n(h)$ as

$$\bar{\boldsymbol{\mu}}_n(h) = -\alpha \boldsymbol{y}_n + (1+\alpha)\hat{\boldsymbol{\mu}}_n(h), \tag{7.2*}$$

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

$$\bar{M}_n(h) = -\alpha I_n + (1+\alpha) M_n(h). \tag{7.3*}$$

It has trace equal to zero. Define

$$\bar{L}_n(h) = n^{-1}\|\boldsymbol{\mu}_n - \bar{\boldsymbol{\mu}}_n(h)\|^2 \quad \text{and} \quad \bar{R}_n(h) = \mathrm{E}\bar{L}_n(h). \tag{7.4*}$$

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

$$n^{-1}\|\boldsymbol{y}_n - \bar{\boldsymbol{\mu}}_n(h)\|^2 + n^{-1} \operatorname{tr} \bar{M}_n(h)\Omega \quad \text{over } \mathscr{H}_n, \tag{7.5*}$$

where

$$\begin{aligned} &n^{-1} \operatorname{tr} \bar{M}_n(h)\Omega \\ &= \left(n^{-1} \operatorname{tr} M_n(h)\Omega - n^{-1} \operatorname{tr} M_n(h) n^{-1} \operatorname{tr} \Omega\right)/\left(1 - n^{-1} \operatorname{tr} M_n(h)\right). \end{aligned} \tag{7.6*}$$

Using Li's (1987) proof of Theorem 3.2, we find that $\bar{h}_G$ is asymptotically optimal for use with the estimator $\hat{\boldsymbol{\mu}}_n(h)$. 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

$$n\bar{R}_n(h) = \Big[\|A_n(h)\boldsymbol{\mu}_n\|^2 + \operatorname{tr} M_n'(h)M_n(h)\Omega$$
$$-2n^{-1}(\operatorname{tr} M_n(h))\operatorname{tr} M_n(h)\Omega + n^{-2}(\operatorname{tr} M_n(h))^2 \operatorname{tr}\Omega\Big]$$
$$\Big/\big(n^{-1}\operatorname{tr} A_n(h)\big)^2. \tag{7.7*}$$

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 $|\operatorname{tr} M_n(h)\Omega|$. 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 $R_n(h) \geq \inf_{i\geq 1}\sigma_i^2 n^{-1}\operatorname{tr} M_n'(h)M_n(h)$. 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

$$L_n(\hat{h}_G)/L_n(\bar{h}_G) \xrightarrow{p} 1. \tag{7.8*}$$

This result plus the asymptotic optimality of $\bar{h}_G$ give the desired result. To show (7.8*), note that

$$n^{-1}\|\boldsymbol{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2/\big(1-n^{-1}\operatorname{tr} M_n(h)\big)^2 = n^{-1}\|\boldsymbol{y}_n - \bar{\boldsymbol{\mu}}_n(h)\|^2. \tag{7.9*}$$

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

$$n^{-1}\|\boldsymbol{y}_n - \hat{\boldsymbol{\mu}}_n(h)\|^2/\big(1-n^{-1}\operatorname{tr} M_n(h)\big)^2 + n^{-1}\operatorname{tr}\bar{M}_n(h)\Omega. \tag{7.10*}$$

In analogy with Li's (1987) argument of (2.1)–(2.4) and in view of (7.6*), (H.1*), and (7.1*), we obtain (7.8*). □

*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*. □

*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 $\sigma^2 h_n n^{-1}$, $\sigma^2 \operatorname{tr} \tilde{M}_n(h_n)\tilde{M}_n'(h_n)$, $\sigma^2(1+o(1)) \operatorname{tr} M_n(h) M_n'(h)$, and $\sigma^2 \tilde{h}_n$ in his proof by $\inf_{i \geq 1} \sigma_i^2 h_n n^{-1}$, $\operatorname{tr} \tilde{M}_n'(h_n) \tilde{M}_n(h_n) \Omega$, $(1+o(1)) \operatorname{tr} M_n'(h) M_n(h) \Omega$, and $\inf_{i \geq 1} \sigma_i^2 \tilde{h}_n$, respectively, and provided one replaces σ^{-4} in his eq. (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 $\tilde{h}_n$ is used as a scalar in Li's proof of Theorem 5.2, it needs to be replaced by $\operatorname{tr} M_n(h_n)$ or $\operatorname{tr} M_n(\tilde{h}_n)$, 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 $\sum_{h \in I_{++}^D} (h'\mathbf{1})^{-m} < \infty$ (as in section 2 above) and the latter holds when $m = D + 1$. □

*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

$$|n^{-1}\|\boldsymbol{\mu}_n - M_n(h)\boldsymbol{\mu}_n\|^2 - n^{-1}\|\boldsymbol{\mu}_n - \tilde{M}_n(h)\boldsymbol{\mu}_n\|^2| \leq 4 w_{n,h}(1)^2 f_\infty^2. \tag{7.11*}$$

In addition, it is straightforward to see that

$$n^{-1} \operatorname{tr} M_n(h) \Omega M_n'(h) = n^{-1} \sum_{i=1}^{n} \sum_{m=1}^{n} \sigma_m^2 \sum_{j=1}^{h} w_{n,h}(j)^2 1(m = i(j)),$$

$$n^{-1} \operatorname{tr} \tilde{M}_n(h) \Omega \tilde{M}_n'(h) = n^{-1} \sum_{i=1}^{n} \sum_{m=1}^{n} \sigma_m^2 \sum_{j=1}^{h+1} w_{n,h}(j-1)^2 1(m = i(j)), \tag{7.12*}$$

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

$$n^{-1} \operatorname{tr} M_n(h) \Omega M_n'(h) \geq \inf_{v \geq 1} \sigma_v^2 \sum_{j=1}^{h} w_{n,h}(j)^2 \tag{7.13*}$$

and

$$|n^{-1}\operatorname{tr} M_n(h)\Omega M_n'(h) - n^{-1}\operatorname{tr}\tilde{M}_n(h)\Omega\tilde{M}_n'(h)|$$

$$\le \sup_{v\ge 1}\sigma_v^2 n^{-1}\sum_{i=1}^{n}\sum_{m=1}^{n}\sum_{j=1}^{h+1} 1(m=i(j))|w_{n,h}(j-1)^2 - w_{n,h}(j)^2|$$

$$= \sup_{v\ge 1}\sigma_v^2 2 w_{n,h}(1)^2, \qquad (7.14^*)$$

where the equality uses (2.8).

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

$$|R_n(h) - \tilde{R}_n(h)| \le \left(4f_\infty^2 + 2\sup_{i\ge 1}\sigma_i^2\right) w_{n,h}(1)^2. \qquad (7.15^*)$$

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

$$w_{n,h}(1)^2 \Big/ \sum_{j=1}^{h} w_{n,h}(j)^2 \xrightarrow{p} 0. \qquad (7.16^*)$$

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).] □

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. and Y.J. Whang, 1990, Additive interactive regression models: Circumvention of the curse of dimensionality, Econometric Theory 6, forthcoming.

Andrews, D.W K, 1991, Asymptotic normality of series estimators for various nonparametric and semiparametric regression models, Econometrica 59, forthcoming.

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, NY).

Cleveland, W.S. and S J Devlin, 1988, Regression by local fitting, Journal of Econometrics 37, 87–114.

Craven, P and G. Wahba, 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 C.R. Hwang, 1982, Nonparametric maximum likelihood estimation by the method of sieves, Annals of Statistics 10, 401–414.

Geweke, J. and R. Meese, 1981, Estimating regression models of finite but unknown order, International Economic Review 22, 55–70.

Härdle, W., P. Hall, and J.S. Marron, 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.

Schmidt, P., 1971, Methods of choosing among alternative models, Workshop paper no. 7004 (Department of Economics, Michigan State University, East Lansing, MI).

Stone, M., 1974, Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society 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, DC) 75–80

Wahba, G., 1990, Spline models in statistics, Regional conference series in applied mathematics (Society for Industrial and Applied Mathematics, Philadelphia, PA).

Wahba, G and S Wold, 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.