

Asymptotic Optimality of Full Cross-validation for Selecting Linear Regression Models

BERND DROGE

*Institute of Mathematics, Humboldt University,
Unter den Linden 6, 10099 Berlin, Germany*

Summary. For the problem of model selection, full cross-validation has been proposed as alternative criterion to the traditional cross-validation, particularly in cases where the latter one is not well defined. To justify the use of the new proposal we show that under some conditions, both criteria share the same asymptotic optimality property when selecting among linear regression models.

AMS 1991 subject classifications: Primary 62J05; secondary 62J99.

Key words: Cross-validation, full cross-validation, model selection, prediction, asymptotic optimality.

1 Introduction

One of the most popular methods for the selection of regression models is based on minimizing the cross-validation (CV) criterion of Stone (1974) among an appropriate class of model candidates. This may be particularly motivated when prediction (or, similarly, estimation of the unknown regression function) is the aim of the statistical analysis. The idea of the traditional leave-one-out CV approach is to assess the

The research on this paper was carried out within the Sonderforschungsbereich 373 at Humboldt University Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft.

predictive performance of a model by an average of certain measures for the ability of predicting one observation by a model fit after deleting just this observation from the data set.

The properties of CV as estimate of the mean squared error of prediction (MSEP) have been compared with those of other model selection criteria by Bunke and Droge (1984). Besides others, there it is shown that some version of bootstrap outperforms CV, which is in accordance with the findings of Efron (1983, 1986). Asymptotic results for the model selection procedure based on CV may be found in Li (1987), where, for example, its asymptotic optimality in the sense of Shibata (1981) is proved.

CV has the appealing feature that no estimation of the error variance is required. On the other hand, there exist nonlinear regression situations where it is not well defined, see e.g. Bunke et al. (1995). To remedy this problem, the so-called full cross-validation (FCV) criterion has been proposed. Its properties as estimate of the MSEP in the linear regression case have been investigated by Droge (1996), with the main result that FCV is superior to CV. However, the conclusions may be different when comparing the behaviour of the model selection procedures based on both criteria, cf. the simulation study in Droge (1995). In the present paper we study the asymptotic behaviour of linear model selection by FCV. It turns out that under some conditions, the minimum-FCV-procedure shares the asymptotic optimality property of the procedure based on CV.

The rest of this paper is organized as follows. The general framework is described in Section 2. There we introduce also the CV and the FCV criteria, whose asymptotic optimality is addressed in Section 3. Section 4 provides a brief discussion of related work. The proof of the main result in Section 3 is deferred to the Appendix.

2 Cross-validation and Full Cross-validation

We assume to have observations y_1, \dots, y_n of a response variable at fixed values x_1, \dots, x_n of a k -dimensional vector of explanatory variables satisfying

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where f is an unknown regression function, and the errors ε_i are independent with mean zero and variance σ^2 . The analysis of the data requires in general to estimate

the function f , for which a variety of parametric and nonparametric approaches exists. In the parametric approach there is seldom sure evidence on the validity of a certain model, so that one has to choose a good one from those being tentatively proposed.

The focus of this paper is on linear model selection. That is, we assume that there are p_n known functions of the explanatory variables, say g_1, \dots, g_{p_n} , associated with the response variable, and the aim is to approximate the regression function by an appropriate linear combination of some of these functions. Each such linear combination is characterized by the subset of indices of the included functions, say $m \subseteq \{1, \dots, p_n\}$. Possibly not all linear combinations are allowed, so that the class of competing models is characterized by a subset M_n of the power set of $\{1, \dots, p_n\}$. Using the least squares approach for fitting the models to the data gives, for each $m \in M_n$, the following estimator of $f(x)$

$$\hat{f}_m(x) = \sum_{i \in m} \hat{\beta}_i^{(m)} g_i(x),$$

where the coefficients $\hat{\beta}_i^{(m)}$ are the minimizers of

$$\sum_{j=1}^n [y_j - \sum_{i \in m} \beta_i g_i(x_j)]^2$$

with respect to β_i ($i \in m$).

On the basis of model m , future values of the response variable at the design point x_i will usually be predicted by $\hat{y}_i(m) = \hat{f}_m(x_i)$ ($i = 1, \dots, n$). Thus, given the observations, the conditional expected squared prediction error is

$$\sigma^2 + L_n(m), \tag{2}$$

where

$$L_n(m) = \frac{1}{n} \sum_{i=1}^n [f(x_i) - f_m(x_i)]^2 \tag{3}$$

is the average squared error loss at the design points. (2) describes the prediction performance of a model, whereas (3) measures the efficiency of model m when estimation of the regression function is the objective of the analysis. Consequently, the prediction problem is closely related to that of estimating f .

Many model selection procedures are based on minimizing criteria which may be interpreted as estimates of (2) or of its unconditional version, the MSE, see e.g. Bunke

and Droge (1984). One of the most widely used MSEP estimates in practice is the CV criterion of Stone (1974) defined by

$$CV(m) = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{y}_{-i}(m)]^2, \quad (4)$$

where $\hat{y}_{-i}(m)$ is the prediction at x_i leaving out the i -th data point. CV works well in many applications. However, to avoid the introductory mentioned difficulties with CV in nonlinear regression, the FCV criterion,

$$FCV(m) = \frac{1}{n} \sum_{i=1}^n [y_i - \tilde{y}_i(m)]^2, \quad (5)$$

has been proposed, where $\tilde{y}_i(m)$ is the least squares prediction at x_i with substituting y_i by $\hat{y}_i(m)$ instead of deleting it, see Bunke et al. (1995) and Droge (1996).

Droge (1996) has compared the different cross-validation criteria as estimates of the MSEP, concluding that FCV outperforms their traditional counterpart. More precisely, it has been shown that the absolute value of the bias of FCV is smaller than that of CV and, under the assumption of normally distributed errors in (1), FCV has also a smaller variance than CV at least in a minimax sense.

3 Asymptotic Optimality

This section is concerned with the asymptotic optimality of the model selection procedures based on minimizing CV and FCV in the sense of Shibata (1981). Let \hat{m} and \tilde{m} denote the minimizer of CV and FCV, respectively, i.e.

$$\hat{m} = \arg \min_{m \in M_n} CV(m) \quad \text{and} \quad \tilde{m} = \arg \min_{m \in M_n} FCV(m).$$

Li (1987) has proved that under reasonable conditions, the minimizer of the CV criterion, \hat{m} , is asymptotically optimal in the sense that, as $n \rightarrow \infty$,

$$\frac{L_n(\hat{m})}{\inf_{m \in M_n} L_n(m)} \rightarrow 1, \quad \text{in probability.} \quad (6)$$

Let $P_n(m)$ be the projection onto the linear space associated with the model indexed by m ("hat matrix"). Then, defining $R_n(m) = EL_n(m)$, the required conditions are the following:

$$(C1) \quad E\varepsilon_1^{4q} < \infty$$

$$(C2) \quad \sum_{m \in M_n} [nR_n(m)]^{-q} \rightarrow 0$$

$$(C3) \quad \inf_{m \in M_n} L_n(m) \xrightarrow{P} 0$$

$$(C4) \quad \lim_{n \rightarrow \infty} \sup_{m \in M_n} \bar{\lambda}(P_n(m)) < 1$$

$$(C5) \quad \exists K > 0 \quad \forall n \quad \forall m \in M_n \quad \bar{\lambda}(P_n(m)) \leq K \frac{|m|}{n}.$$

Here, conditions (C1) and (C2) are assumed to hold for some natural number q , $\bar{\lambda}(\cdot)$ denotes the maximum diagonal element of a matrix and $|m|$ is the dimension of the model (number of elements in m).

For the minimizer of the FCV criterion, \tilde{m} , the same property may be shown under the additional assumption that the largest model dimension increases slower than the sample size.

Theorem. *Assume that (C1), (C2) and $c_n := \sup_{m \in M_n} \bar{\lambda}(P_n(m)) = o(1)$ hold. Then \tilde{m} is asymptotically optimal, i.e. (6) holds with \tilde{m} instead of \hat{m} .*

Remarks on the assumptions. Assume that the explanatory variables, say $x_{(1)}, \dots, x_{(p_n)}$, are given in a decreasing order of importance such as in polynomial regression. Then it is quite natural to consider only the case of nested models, i.e. $M_n = \{\{1\}, \{1, 2\}, \dots, \{1, 2, \dots, p_n\}\} =: M_n^*$, where each model $m \in M_n^*$ is identified by the set of indices of those explanatory variables which are included in the model. For this situation, Li (1987) has shown that only $q = 2$ is needed for the moment condition (C1) and, moreover, (C2) may be replaced by the weaker condition

$$(C2') \quad \inf_{m \in M_n} nR_n(m) \rightarrow \infty, \text{ as } n \rightarrow \infty.$$

For example, from Shibata (1981) it is known that in the problem of selecting an appropriate order in polynomial regression, the condition (C2') will hold when the true regression function is not a polynomial.

As noticed by Li (1987), condition (C3) assumes only the existence of a consistent selection procedure when f is known.

Condition (C4) requires that the diagonal elements of the hat matrix are bounded away from 1. Recalling that $P_n(m)$ is a projection matrix leading to $\bar{\lambda}(P_n(m)) \leq 1$, this may be recognized as a weak assumption. ($p_{ii}(m) = 1$ implies that the vector of the regression parameters is not identifiable when leaving out the i -th data point.)

Condition (C5) excludes extremely unbalanced designs, see again Li (1987).

Finally we mention that our condition $c_n \rightarrow 0$ follows e.g. from (C5) if the largest

model dimension increases slower than the sample size, i.e. $\sup_{m \in M_n} |m| = o(n)$. In the case of nested models this corresponds just to the assumption $p_n = o(n)$, which was also imposed by Shibata (1981). However, such a condition makes the selection rule not completely data-driven, since it is hard to decide whether $\sup_{m \in M_n} |m|$ is small enough compared with n .

4 Some Related Work

The asymptotic properties of model selection procedures based on different criteria have been investigated by several authors. Here we present only a short review of some results.

Nishii (1984) considered the problem of selecting an appropriate submodel of some given linear model, say $m_1 = \{1, \dots, p\}$ with associated design matrix G , of fixed dimension $p_n = p$. Consequently, $M_n = M$ does not depend on the sample size. He made the following assumption:

(N) There is a true (minimal adequate) linear regression model, say $m_0 \in M$. The matrix $G^T G$ is positive definite, and $\lim_{n \rightarrow \infty} n^{-1} G^T G$ exists and is also positive definite.

We will reformulate Nishii's result in terms of two notions of consistency for a model selection procedure \bar{m} , which have been introduced by Müller (1993). A procedure \bar{m} is called *m_0 -consistent* if its probability of selecting the true model tends to one, i.e. $P(\bar{m} = m_0) \rightarrow 1$ as $n \rightarrow \infty$. Moreover, with $M_0 = \{m \in M_n \mid m_0 \subseteq m\}$, \bar{m} is called *M_0 -consistent* if $P(\bar{m} \in M_0) \rightarrow 1$ as $n \rightarrow \infty$, i.e. if the probability of selecting a model not including the true one tends to zero. Then Nishii showed that under (N) and the assumption of normally distributed errors, procedures based on criteria CV, C_p (Mallows, 1973), FPE (Akaike, 1970) and AIC (Akaike, 1974) are M_0 -consistent but not m_0 -consistent, that is the selected models apt to overfit. In contrast, the m_0 -consistency was proved for the criterion GIC which is a generalization of BIC (Schwarz, 1978). Note that for the result on CV, $\lim_{n \rightarrow \infty} \bar{\lambda}(P_n(m_1)) = 0$ is additionally required, which is in this case equivalent to our condition $c_n \rightarrow 0$.

The above results have been generalized by Müller (1993) to the case of nonnormal errors and inadequate linear models defining a pseudo-true (instead of a true) model in a convenient way and assuming some additional conditions on the design and the

unknown regression function. Now it is easy to check that under the same assumptions the minimum-FCV-procedure \tilde{m} is also M_0 -consistent. As remarked by Müller (1993), the results can be generalized to cases where the dimension of model m_1 increases with the sample size, i.e. for $p = p_n = o(n)$, but that of the true model m_0 is still fixed. Furthermore, assuming certain conditions to ensure that m_0 minimizes $L_n(m)$ for sufficiently large n , which are fulfilled e.g. under (N), m_0 -consistent procedures may be seen to be also asymptotically optimal in the sense of (6).

In the situation of Nishii (1984) but with errors as in (1), Shao (1993) made similar observations concerning the asymptotic behaviour of the minimum-CV-procedure. He found that the deficiency of the leave-one-out CV can be rectified by using a leave- d -out CV, say $CV(d)$. More precisely, he showed that some variants of $CV(d)$ are m_0 -consistent if $d/n \rightarrow 1$ and $n - d \rightarrow \infty$ as $n \rightarrow \infty$ and, with the notation $\mu = (f(x_1), \dots, f(x_n))^T$, the following conditions are satisfied:

$$\begin{aligned} \liminf_{n \rightarrow \infty} n^{-1} \|(I - P_n(m))\mu\|^2 &> 0 \text{ for models } m \in M \setminus M_0 & (7) \\ G^T G = O(n), \quad (G^T G)^{-1} = O(n^{-1}), \quad c_n = o(1). \end{aligned}$$

Notice that condition (7) on the model biases provides some type of asymptotic model identifiability.

Zhang(1993) dealt with multifold CV in the same context, too. Under some assumptions including $d/n \rightarrow \delta > 0$ as $n \rightarrow \infty$ and $c_n = o(1)$, he established that the $CV(d)$ criterion is asymptotically equivalent to the criterion

$$CR(\alpha, m) = RSS(m) + \alpha |m| \hat{\sigma}^2(m_1), \quad \text{with } \alpha = (2 - \delta)/(1 - \delta), \quad (8)$$

where $RSS(m) = \sum_{i=1}^n [y_i - \hat{y}_i(m)]^2$ is the residual sum of squares under model m , and $\hat{\sigma}^2(m_1) = RSS(m_1)/(n - |m_1|)$. Obviously, it holds $\alpha > 2$ if $\delta > 0$, whereas $CR(2, m)$ may be recognized by the reader as the C_p -criterion of Mallows (1973). Furthermore, Zhang's results imply that under his assumptions the $CV(d)$ -method is M_0 -consistent but not m_0 -consistent. This is in some accordance with the above result of Shao (1993), who proved the necessity of $d/n \rightarrow 1$ for m_0 -consistency, although this condition seems rather surprising at first glance. Another interesting conclusion of Zhang is that the probability of choosing the true model m_0 is an increasing function of δ . When $\delta \rightarrow 0$, the $CV(d)$ criterion becomes equivalent to the CV criterion.

Model selection procedures based on minimizing the criterion (8) were also investigated by Zheng and Loh (1995), assuming that the "covariates" g_i are either preordered

or sorted according to t -statistics. Thus, the competing models are nested as in M_n^* . The errors in (1) were assumed to be sub-Gaussian and, moreover, the maximal model dimension p_n was allowed to depend on n , satisfying $\limsup_n p_n/n < 1$, whereas the true model did not depend on the sample size. The authors showed how the factor $\alpha|m|$ of the penalty term for the model complexity in (8) has to be replaced by some positive nondecreasing function, $h_n(|m|)$, of $|m|$ to achieve m_0 -consistency of the corresponding model selection procedure. The imposed condition on the design matrices is fairly minimal in proving asymptotic theory for linear models (and weaker than that of Shao, 1993), whereas the growth restrictions on h_n reveal the interplay of h_n , p_n and the minimal bias of an inadequate model, $\Gamma_{min} = \min_{m \in M_n \setminus M_0} \|(I - P_n(m))\mu\|$, which is necessary for preventing both overfitting and underfitting. The m_0 -consistency of a procedure depends clearly on the choice of h_n , which in turn is decided by p_n and the growth of Γ_{min} since, roughly speaking, h_n has to increase faster than p_n but slower than Γ_{min} . Generally, if $p_n \rightarrow \infty$ as $n \rightarrow \infty$, then the penalty $h_n(|m|)$ is required to grow faster than when p_n is bounded. Under the imposed assumptions it turns out that, for example, BIC (defined by $h_n(|m|) = |m| \log n$) is m_0 -consistent if $p_n = o(\log n)$. On the other hand, $p_n = o(\log \log n)$ would imply the m_0 -consistency of the ϕ criterion given by $h_n(|m|) = c|m| \log \log n$, where $c > 0$ (Hannan and Quinn, 1979).

It should be pointed out that all results on which we have commented above depend heavily on the assumed existence of a fixed finite-dimensional true (or pseudo-true) model. The story is quite different when the dimension of the true model increases with the sample size or is infinite. In this case it is already known from Shibata (1981) that criteria with comparatively small penalties for the model complexity such as AIC, FPE and C_p are optimal in the sense of (6), whereas those with larger penalties like BIC and ϕ are not. The results of the present paper on FCV as well as those of Li (1987) on CV, C_p and generalized CV of Craven and Wahba (1979) are in the same spirit. We remark that Li (1987) treated the somewhat more general problem of selecting a good estimate from a proposed class of linear estimates indexed by some discrete set, covering, for instance, also the nearest-neighbour nonparametric regression case.

Appendix: Proof of Theorem

Let $P_n(m) = ((p_{ij}(m)))_{i,j=1,\dots,n}$, $y = (y_1, \dots, y_n)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, and $\|z\|_A = z^T A z$ for an $n \times n$ -matrix A and $z \in \mathbb{R}^n$. Then we derive from (3.8) and (3.10) in Droge (1996) that

$$FCV(m) = n^{-1} \|\varepsilon\|^2 + L_n(m) + Z_n(m),$$

where

$$Z_n(m) = \frac{2}{n}(\mu - P_n(m)y)^T \varepsilon + \frac{1}{n} \|y - P_n(m)y\|_{\Delta(m)}^2, \quad \text{and}$$

$$\Delta(m) = \text{diag}[\delta_1(m), \dots, \delta_n(m)], \quad \delta_i(m) = p_{ii}(m)(2 + p_{ii}(m)).$$

As we will see, it suffices to verify that $Z_n(m)$ is negligible (compared with $L_n(m)$) uniformly for any $m \in M_n$. More precisely, we will show that in probability,

$$\sup_{m \in M_n} |Z_n(m)|/R_n(m) \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad (\text{A1})$$

leading immediately to

$$\sup_{m, m' \in M_n} \frac{|L_n(m) - L_n(m') - [FCV(m) - FCV(m')]|}{L_n(m) + L_n(m')} \xrightarrow{P} 0, \quad (\text{A2})$$

since the expression on the left-hand side is bounded from above by $2 \sup_{m \in M_n} |Z_n(m)|/L_n(m)$ and, as established by Li (1987) under the conditions (C1) and (C2),

$$\sup_{m \in M_n} |L_n(m)/R_n(m) - 1| \xrightarrow{P} 0.$$

The desired result is now a consequence of (A2): Given any $\eta > 0$, set $\gamma = \eta/(2 + \eta)$. Then, definig $m^* = \arg \min_{m \in M_n} L_n(m)$ and recalling the definition of \tilde{m} , we obtain

$$\begin{aligned} P \left\{ \left| \frac{L_n(\tilde{m})}{L_n(m^*)} - 1 \right| > \eta \right\} &= P \left\{ \frac{L_n(\tilde{m})}{L_n(m^*)} > \frac{1 + \gamma}{1 - \gamma} \right\} \\ &= P \{ (1 - \gamma)L_n(\tilde{m}) - (1 + \gamma)L_n(m^*) > 0 \} \\ &\leq P \{ (1 - \gamma)L_n(\tilde{m}) - (1 + \gamma)L_n(m^*) > FCV(\tilde{m}) - FCV(m^*) \} \\ &\leq P \{ |L_n(\tilde{m}) - L_n(m^*) - [FCV(\tilde{m}) - FCV(m^*)]| > \gamma[L_n(\tilde{m}) + L_n(m^*)] \} \\ &\leq P \left\{ \sup_{m, m' \in M_n} \frac{|L_n(m) - L_n(m') - [FCV(m) - FCV(m')]|}{L_n(m) + L_n(m')} > \gamma \right\}, \end{aligned}$$

which converges to zero due to (A2).

Thus it remains to show (A1). To accomplish this, we use the decomposition

$$Z_n(m) = S_1(m) + 2S_2(m) + S_3(m),$$

where

$$\begin{aligned} S_1(m) &= \frac{1}{n} \|(I - P_n(m))\mu\|_{\Delta(m)}^2, \\ S_2(m) &= \frac{1}{n} \mu^T [(I - P_n(m))\Delta(m)(I - P_n(m)) + I - P_n(m)]\varepsilon, \quad \text{and} \\ S_3(m) &= \frac{1}{n} \varepsilon^T [(I - P_n(m))\Delta(m)(I - P_n(m)) - 2P_n(m)]\varepsilon \end{aligned}$$

and establish that, in probability, $\sup_{m \in M_n} |S_i(m)|/R_n(m) \rightarrow 0$ for $i = 1, 2, 3$.

(i) To prove the statement for $S_1(m)$, we observe first

$$R_n(m) = \frac{1}{n} \|\mu - P_n(m)\mu\|^2 + \frac{\sigma^2}{n} \text{tr}[P_n(m)]. \quad (\text{A3})$$

Obviously we have $\delta_i(m) \leq c_n(2 + c_n)$ for $i = 1, \dots, n$, which implies on account of (A3), for all $m \in M_n$,

$$S_1(m) = \frac{1}{n} \|(I - P_n(m))\mu\|_{\Delta(m)}^2 \leq \frac{c_n(2 + c_n)}{n} \|(I - P_n(m))\mu\|^2 \leq c_n(2 + c_n)R_n(m), \quad (\text{A4})$$

and hence $S_1(m)/R_n(m) \leq c_n(2 + c_n)$. The desired result follows since $c_n \rightarrow 0$ as $n \rightarrow \infty$.

(ii) Given any $\eta > 0$, we have for some constant $C > 0$

$$\begin{aligned} P \left\{ \sup_{m \in M_n} \frac{|S_2(m)|}{R_n(m)} > \eta \right\} &\leq \sum_{m \in M_n} P \left\{ \frac{|S_2(m)|}{R_n(m)} > \eta \right\} \\ &\leq \sum_{m \in M_n} \frac{E|\mu^T (I - P_n(m))(I + \Delta(m))(I - P_n(m))\varepsilon|^{2q}}{n^{2q}\eta^{2q}R_n(m)^{2q}} \quad (\text{A5}) \end{aligned}$$

$$\leq C\eta^{-2q} \sum_{m \in M_n} \frac{\|(I - P_n(m))(I + \Delta(m))(I - P_n(m))\mu\|^{2q}}{n^{2q}R_n(m)^{2q}} \quad (\text{A6})$$

$$\leq C(1 + c_n(2 + c_n))^{2q}\eta^{-2q} \sum_{m \in M_n} [nR_n(m)]^{-q}, \quad (\text{A7})$$

which tends zero as $n \rightarrow \infty$ due to condition (C2). Notice that (A5) and (A6) follow because of the Markov inequality and Theorem 2 of Whittle (1960), respectively, whereas (A7) may be derived similarly to (A4).

(iii) With the notation $H(m) = (I - P_n(m))\Delta(m)(I - P_n(m)) - 2P_n(m)$, the last term may be rewritten as $S_3(m) = n^{-1}\|\varepsilon\|_{H(m)}^2$. Recalling $P_n^2(m) = P_n(m)$,

$\delta_i(m) = p_{ii}(m)(2 + p_{ii}(m)) \leq c_n(2 + c_n)$ and (A3), it is easily seen that there is some constant $K > 0$ such that

$$\begin{aligned} \text{tr}[H(m)H^T(m)] &= \text{tr}[\Delta(m)(I - P_n(m))]^2 + 4\text{tr}[P_n(m)] \\ &\leq K\text{tr}[P_n(m)] \leq Kn\sigma^{-2}R_n(m). \end{aligned} \quad (\text{A8})$$

Given any $\eta > 0$, we conclude therefore by the same arguments as in (ii) that, for some $C^* > 0$,

$$\begin{aligned} P \left\{ \sup_{m \in M_n} \frac{|S_3(m) - ES_3(m)|}{R_n(m)} > \eta \right\} &\leq \sum_{m \in M_n} \frac{E|S_3(m) - ES_3(m)|^{2q}}{n^{2q}\delta^{2q}R_n(m)^{2q}} \\ &\leq \eta^{-2q}C^* \sum_{m \in M_n} \frac{\{\text{tr}[H(m)H^T(m)]\}^q}{[nR_n(m)]^{2q}} \\ &\leq \eta^{-2q}\sigma^{-2q}C^*K^q \sum_{m \in M_n} [nR_n(m)]^{-q}, \end{aligned}$$

which, again on account of (C2), converges to zero as $n \rightarrow \infty$. Hence the proof is completed by showing that, uniformly for any $m \in M_n$, $|ES_3(m)|/R_n(m) \rightarrow 0$ as $n \rightarrow \infty$. To accomplish this we notice that

$$\begin{aligned} ES_3(m) &= \frac{1}{n}E\|\varepsilon\|_{H(m)}^2 = \frac{\sigma^2}{n}\text{tr}[H(m)] \\ &= \frac{\sigma^2}{n}\text{tr}[\Delta(m) - \Delta(m)P_n(m) - 2P_n(m)] = -\frac{\sigma^2}{n} \sum_{i=1}^n p_{ii}^2(m)(1 + p_{ii}(m)). \end{aligned}$$

Consequently, we have

$$|ES_3(m)|/R_n(m) \leq c_n(1 + c_n)n^{-1}\sigma^2\text{tr}[P_n(m)] \leq c_n(1 + c_n)R_n(m),$$

which in turn entails the desired result since $c_n \rightarrow 0$ as $n \rightarrow \infty$.

References

- AKAIKE, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203-217.
- AKAIKE, H. (1974). A new look at the statistical model identification. *I.E.E.E. Trans. Auto. Control* **19**, 716-723.

- BUNKE, O. and DROGE, B. (1984). Bootstrap and cross-validation estimates of the prediction error for linear regression models. *Ann. Statist.* **12**, 1400-1424.
- BUNKE, O., DROGE, B. and POLZEHL, J. (1995). Model selection, transformations and variance estimation in nonlinear regression. Discussion Paper No. 95-52, Sonderforschungsbereich 373, Humboldt-Universität, Berlin.
- 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. *Numer. Math.* **31** 377-403.
- DROGE, B. (1995). Some simulation results on cross-validation and competitors for model choice. In: *MODA4 – Advances in Model Oriented Data-Analysis* (Eds. C.P. Kitsos and W.G. Müller), Physica, Heidelberg, 213-222.
- DROGE, B. (1996). Some comments on cross-validation. In: *Statistical Theory and Computational Aspects of Smoothing* (Eds. W. Härdle and M.G. Schimek), Physica, Heidelberg, 178-199.
- EFRON, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316-331.
- EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461-470.
- HANNAN, E.J. and QUINN, B.G. (1979). The determination of the order of autoregression. *J. Roy. Statist. Soc.* **B41**, 190-195.
- LI, K.-C. (1987). Asymptotic optimality for C_p , C_L , cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15**, 958-975.
- MALLOWS, C.L. (1973). Some comments on C_p . *Technometrics* **15**, 661-675.
- MÜLLER, M. (1993). Asymptotische Eigenschaften von Modellwahlverfahren in der Regressionsanalyse. Doctoral Thesis, Department of Mathematics, Humboldt University, Berlin (in German).
- NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 758-765.

- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.
- SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.
- SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.
- STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. B***36**, 111-147.
- WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5**, 302-305.
- ZHANG, P. (1993). Model selection via multifold cross-validation. *Ann. Statist.* **21**, 299-313.
- ZHENG, X. and LOH, W.-Y. (1995). Consistent variable selection in linear models. *J. Amer. Statist. Assoc.* **90**, 151-156.