Droge, Bernd

Working Paper
Asymptotic properties of model selection procedures in linear regression

SFB 373 Discussion Paper, No. 2003,28

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin


This Version is available at:
http://hdl.handle.net/10419/22243

Terms of use:
Documents in EconStor may be saved and copied for your personal and scholarly purposes.
You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.
Asymptotic Properties of Model Selection Procedures in Linear Regression

BERND DROGE

SFB 373, Humboldt University,
Unter den Linden 6, 10099 Berlin, Germany

Summary. In regression analysis there is typically a large collection of competing models available from which we want to select an appropriate one. This paper is concerned with asymptotic properties of procedures for selecting linear models, which are based on certain data-dependent criteria such as Mallows’ $C_p$, cross-validation and the generalized information criterion. We avoid the assumption of an adequate (“correct”) model and allow the maximal model dimension to increase with the sample size. General asymptotic concepts are introduced, covering the usual ones of consistency and asymptotic optimality. The focus is on conditions for penalizing the model complexity which are necessary to obtain the different optimalities. For example, the consistency of a procedure is decided by the interplay between these penalties, the complexity of the class of model candidates, and some quantity describing the ability to identify “wrong” (pseudo-inadequate) models. Many results known from the literature appear as special cases or are slightly modified.


Key words: Model selection, prediction, asymptotic optimality, consistency.

1 Introduction

We assume to have observations $y_1, \ldots, y_n$ of a response variable (with values in $\mathcal{Y}$) at fixed values $x_1, \ldots, x_n$ of a $k$-dimensional vector of explanatory variables satisfying

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$  \hspace{1cm} (1.1)
where \( f \) is an unknown regression function, and the errors \( \varepsilon_i \) are independent with mean zero and variance \( \sigma^2 > 0 \). 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 \( p_n, p_n \leq n \), known functions of the explanatory variables, say \( g_1, \ldots, g_{p_n} \), 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 \( m \subseteq \{1, \ldots, p_n\} =: m_1 \). 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 \( m_1 \). 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 \( \beta_i (i \in m) \). Note that \( m = \emptyset \) may also be considered as model candidate, leading to \( \hat{f}_\emptyset \equiv 0 \).

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

\[
\sigma^2 + L_n(m),
\]

(1.2)

where

\[
L_n(m) = \frac{1}{n} \sum_{i=1}^{n} [f(x_i) - \hat{f}_m(x_i)]^2
\]

(1.3)
denotes the average squared error loss at the design points. (1.2) describes the prediction performance of a model, whereas (1.3) 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 \( f \).
With $\mu = (f(x_1), \ldots, f(x_n))^T$, $\hat{\mu}_m = (\hat{f}_m(x_1), \ldots, \hat{f}_m(x_n))^T$ and the Euclidean norm $\|\cdot\|$, (1.3) may be rewritten as

$$L_n(m) = \frac{1}{n}\|\mu - \hat{\mu}_m\|^2.$$  

The risk associated with this loss is then the mean squared error for estimating $f$, $R_n(m) = E L_n(m)$. Using the notation

$$G_m = ((g_j(x_i))_{i=1}^{n})$, $G = G_{m_1}$

and

$$P_m = G_m(G_m^T G_m)^{-1} G_m^T = ((p_{ij}(m)))_{i,j=1}^{n}$, $P = P_{m_1}$

the risk may be decomposed into the so-called model error (or model bias) and the estimation error:

$$R_n(m) = \Delta_n(m) + \sigma^2 n^{-1} |m|,$$  \hspace{1cm} (1.4)

where

$$\Delta_n(m) = \min_{\beta_m \in R} \frac{1}{n} \|\mu - G_m \beta_m\|^2 = \frac{1}{n} \|\mu - P_m \mu\|^2,$$

and $|m| = \text{tr}[P_m]$ denotes the dimension of the model (number of elements in $m$). Throughout this paper we assume that the design matrix, $G$, of the largest possible model, $m_1$, has full rank.

If the aim of the statistical analysis is prediction (or, similarly, the estimation of the unknown regression function), one would ideally select a model by minimizing (1.2) or its unconditional version, the mean squared error of prediction (MSEP) defined by $MSEP(m) = \sigma^2 + R_n(m)$, among the class $M_n$ of model candidates. Since the MSEP is unknown, it seems to be reasonable that many model selection procedures are based on minimizing some criterion which may be interpreted as estimate of the MSEP or of some transformation of it, compare e.g. Bunke and Droge (1984a). In what follows we present some criteria which are in common use for model selection.

One of the most widely used MSEP estimates in practice is the cross-validation (CV) criterion of Stone (1974) defined by

$$CV(m) = \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{y}_{-i}(m)]^2,$$  \hspace{1cm} (1.5)

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, it may fail in some nonlinear regression situations, as
it has been illustrated in Bunke et al. (1999). To avoid these difficulties with CV in nonlinear regression, the full cross-validation (FCV) criterion,

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

(1.6)

has been proposed, where \( \hat{y}_i(m) \) 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. (1999) and Droge (1996).

Craven and Wahba (1979) have proposed generalized cross-validation (GCV) as another useful method for selecting the smoothing parameter of linear estimates, which may also be applied for model selection. With \( y = (y_1, \ldots, y_n)^T \), let

\[
RSS(m) = \frac{1}{n} \| y - \hat{\mu}_m \|^2
\]

be the residual sum of squares under model \( m \). Then, assuming that \( |m| < n \), the GCV criterion is defined by

\[
GCV(m) = RSS(m)/(1 - n^{-1}|m|^2).
\]

It may be seen that GCV weights the ordinary residuals \( (y_i - \hat{y}_i(m)) \) by the average of the weights used for the CV criterion. Applying this idea to FCV, Droge (1996) has introduced the following generalized full cross-validation (GFCV) criterion

\[
GFCV(m) = RSS(m)(1 + n^{-1}|m|^2).
\]

The above CV, GCV, FCV and GFCV criteria do not require estimation of the error variance. This is an advantage over other model selection criteria such as Mallows’ (1973) \( C_p \), which is given by

\[
C_p(m) = RSS(m) + 2n^{-1}|m|\hat{\sigma}^2,
\]

(1.7)

where \( \hat{\sigma}^2 \) denotes some appropriate estimate of \( \sigma^2 \). Other frequently discussed criteria are the following:

\[
FPE(m) = \frac{n + |m|}{n - |m|} RSS(m) \quad \text{(final prediction error, Akaike, 1970),}
\]

\[
SH(m) = \frac{n + 2|m|}{n} RSS(m) \quad \text{(Shibata, 1981),}
\]

\[
GIC(m) = \ln(RSS(m)) + a_n n^{-1}|m| \quad \text{(Nishii, 1984),}
\]

where \( a_n > 0 \) is a sequence with \( a_n = o(n) \). We remark that the generalized information criterion, GIC, covers several other well known criteria as special cases under the
assumption of normally distributed observations. We obtain, for example, Akaike’s information criterion (AIC, Akaike, 1974) for $a_n = 2$, the Bayesian information criterion (BIC, Schwarz, 1978) for $a_n = \ln(n)$ and the criterion $\phi$ of Hannan and Quinn (1979) for $a_n = 2 \ln(\ln(n))$.

Bunke and Droge (1984a,b) and Droge (1996) have compared the different criteria as estimates of the MSEP. Besides others, Bunke and Droge (1984a) have shown that some version of the bootstrap outperforms CV, which is in accordance with the findings of Efron (1983, 1986). Moreover, this bootstrap criterion turns out to be equivalent to the $C_p$-criterion in important special cases. Droge (1996) has particularly investigated the different cross-validation criteria, concluding that FCV and GFCV outperform their traditional counterparts. More precisely, it has been shown that the absolute value of the biases of FCV and GFCV are smaller than those of CV and GCV, respectively. Moreover, under the assumption of normally distributed errors in (1.1) it holds $\text{Var}(GFCV) < \text{Var}(GCV)$, and FCV has also a smaller variance than CV at least in a minimax sense.

In assessing the results of analyzing experimental data we have to take into account the fact that the obtained predictions and estimates may depend heavily on the chosen model. As commented above, in practice this model is often selected by employing data-driven automated methods. Although the use of most criteria may be motivated by being estimates of the MSEP, it is thus not clear that a better MSEP estimate provides a more appropriate model selection criterion in the sense of leading to a model with smaller overall MSEP; compare e.g. the simulation study in Droge (1995). However, it is hard to establish finite-sample properties of model selection procedures, and consequently most results in this field are asymptotic in character. Therefore, the asymptotic behaviour of such procedures is addressed in this paper; for some small sample results within a decision-theoretic framework we refer to Droge (1993) and Droge and Georg (1995).

In the literature there exist mainly two notions for characterizing the asymptotic behaviour of model selection procedures: consistency and asymptotic optimality. These notions are introduced in a general context in Section 2. There we comment also on the relations between both notions and give an overview about the corresponding asymptotic properties which may be established in various situations for so-called canonical model selection procedures. Such procedures assume a known error variance and are characterized by nonrandom penalties for the model complexity. We elaborate on the interplay between these penalties and other quantities, which is necessary to achieve
the different optimalities. Giving up the assumption of a known variance, the results are then applied in Section 3 to derive asymptotic properties of a variety of data-dependent model selection procedures which are in common use. We are particularly concerned with all criteria introduced above. In this way, many results known from the literature are covered and partly generalized. Finally, Section 4 provides a brief discussion of related work. All proofs are deferred to an appendix.

2 Asymptotic properties of canonical procedures

2.1 Notions and preliminary results

In general, none of the models in $M_n$ will be correct, that is there don’t exist coefficients $\beta_i^0$ with $f = \sum_{i=1}^{p_n} \beta_i^0 g_i$. Interest is then in estimating the projection parameter

$$\beta^f = (\beta_1^f, \ldots, \beta_{p_n}^f)^T = \arg \min_{\beta} \| G\beta - \mu \|^2,$$  

(2.1)

which is unique since $G$ was assumed to be nonsingular. In this case, we can define a pseudo-true model by

$$m_0 = \{i \in m_1 \mid \beta_i^f \neq 0\}.$$

Obviously, both $\beta^f$ and $m_0$ may depend on $n$ (and $m_1$), but for the sake of simplicity we have not indicated this in the notation. Furthermore, the full rank property of $G$ gives immediately that $m_0$ is the minimizer of the model bias with the smallest dimension:

**Lemma 2.1** Let $m_0$ be the pseudo-true model and $\delta_n = \min_{m \in M_n} \Delta_n(m)$. Then we have $\Delta_n(m_0) = \delta_n$ and, for any minimizer $\tilde{m}$ of $\Delta_n(m)$, $m_0 \subseteq \tilde{m}$ and $P_{m_0}\mu = P_m\mu$.

Let $M_0 = \{m \in M_n \mid \Delta_n(m) = \delta_n\}$ be the set of pseudo-adequate models, i.e. of all models with minimal model bias. Note that Lemma 2.1 and its proof provide $M_0 = \{m \in M_n \mid m_0 \subseteq m\}$. In the case that there is a correct model, we have $\delta_n = 0$ and thus $\Delta_n(m) = 0$ for all $m \in M_0$. Then $m_0$ and $M_0$ will be called true model and set of adequate models, respectively.

Let $D_n(m)$ be some (positive) functional (loss or discrepancy measure) describing the accuracy of a given model $m$, i.e. of the estimator of $f$ based on $m$. An appropriate model is always selected from the class of all admitted models $M_n$. However, the user may have interest only in models possessing certain properties such as pseudo-adequacy, which form some (possibly unknown) subset $M_n^0 \subseteq M_n$. Ideally, one would then try to minimize $D_n(m)$ over $m \in M_n^0$. But this is in general impossible, since $D_n(m)$
depends typically on unknown quantities like \( f \) and \( \sigma \). Nevertheless, the aim is to achieve the optimality at least asymptotically. Therefore, the following notions are usually considered in the context of an asymptotic theory.

**Definition.** Let \( M^*_n = M^*_n(D_n, M_n^0) = \arg\min_{m \in M_n^0} D_n(m) \) be the set of all models minimizing the discrepancy \( D_n(m) \) over \( M_n^0 \) and \( d_n = d_n(D_n, M_n^0) = \min_{m \in M_n^0} D_n(m) \) denote the associated value of the discrepancy, which is assumed to be positive (almost surely). Then a model selection procedure \( \hat{m} : \mathcal{Y} \rightarrow M_n \) is called \( M^*_n \)-consistent if \( P(\hat{m} \in M_n^*) \rightarrow 1 \) as \( n \rightarrow \infty \). The procedure \( \hat{m} \) is called asymptotically \( d_n \)-optimal if

\[
\frac{D_n(\hat{m})}{d_n} \xrightarrow{P} 1 \quad \text{as} \quad n \rightarrow \infty.
\]

Note that the set of optimal models \( M^*_n \) may consist of more than one model and may be random for random discrepancy measures. In the following we will shortly use the notions of consistent and asymptotically optimal model selection procedures, respectively, if the underlying discrepancy measure and the set \( M_n^0 \) are clear from the context. Consequently, for a consistent model selection procedure the probability of selecting an “optimal” model tends to one as \( n \rightarrow \infty \). Moreover, using an asymptotically optimal procedure, one does asymptotically as well as if one knew the true regression function, provided one restricts to the use of the (linear) least squares estimators \( \hat{f}_m \).

The following straightforward fact shows the relations between the introduced notions.

**Proposition 2.1.**

(i) Any consistent model selection procedure is asymptotically optimal.

(ii) Conversely, an asymptotically optimal procedure \( \hat{m} \) is also consistent, if, as \( n \rightarrow \infty \), \( P(\hat{m} \in M_n^0) \rightarrow 1 \) and, with \( d_n^* = d_n^*(D_n, M_n^0) = \min_{m \in M_n^0 \setminus M_n^*} D_n(m) - d_n \),

\[
d_n/d_n^* = O_P(1). \tag{2.2}
\]

Note that (2.2) may be seen as (asymptotic) identifiability condition for the (set of) optimal model(s), and is not a condition on the model selection procedure.

By some examples we illustrate now the situation for different specifications of \( D_n \) and \( M_n^0 \).

**Example 1.** Using \( D_n(m) = R_n(m) \) and \( M_n^0 = M_0 \) we obtain \( M_n^* = \{m_0\} \) and \( d_n = \delta_n + n^{-1}|m_0|\sigma^2 \). That is, minimizing the risk over all pseudo-adequate models provides the (unique) pseudo-true model, and the consistency considerations
correspond to the usual consistency approach in the literature (see e.g. Nishii, 1984, for the adequate case or Müller, 1993, for the inadequate case but with an asymptotically true model instead of a pseudo-true model depending on $n$), which we will refer to as $m_0$-consistency. Selecting the pseudo-true model is of particular interest, for example, in pilot studies to larger experiments, where it is not of interest that the potentially best fit (providing the minimal model bias) can be achieved with the given sample, see e.g. Linhart & Zucchini (1986).

**Example 2.** Choosing $D_n(m) = \Delta_n(m)$ and $M_n^0 = M_n$ leads to $d_n = \delta_n$ and $M_n^* = M_0$, so that the class of all pseudo-adequate models would be the focus of the investigation. The corresponding consistency approach is known as $M_0$-consistency (Müller, 1993). This property is useful for proving the consistency results in the context of Example 1 (cp. Proposition 2.2). In this setting it is easy to verify that

$$
\delta_n^* := d_n^*(\Delta_n, M_n) = \min_{m \in M_n \setminus M_0} \Delta_n(m) - \delta_n = \min_{m \in M_n \setminus M_0} \frac{1}{n} \| (P - P_m) \mu \|^2 , \quad (2.3)
$$

compare Proposition 2.1 for the definition of $d_n^*$. Asymptotic optimality considerations are irrelevant to this case, since the existence of an adequate (true) model would imply $\delta_n = 0$. □

**Example 3.** If one is interested in the behaviour of a procedure for the sample at hand, one could consider $D_n(m) = L_n(m)$ and $M_n^0 = M_n$. Then the minimizer of the loss $L_n(m)$ over all admitted models is not necessarily unique. The optimal (random) set $M_n^* := M_n^*(L_n, M_n)$ corresponds to the theoretically optimal choice for the sample at hand. In this situation, our asymptotic optimality notion corresponds to that of Li (1987), and we will refer to it as asymptotic $l_n$-optimality, where

$$
l_n := d_n(L_n, M_n) = \min_{m \in M_n} L_n(m) ;
$$

see also Shibata (1981) for a similar approach. □

**Example 4.** If one would like to find a model which minimizes the risk instead of the loss, i.e. if $D_n(m) = R_n(m)$ and $M_n^0 = M_n$, then $M_n^* := M_n^*(R_n, M_n)$ as well as

$$
r_n := d_n(R_n, M_n) = \min_{m \in M_n} R_n(m)
$$

are nonrandom. However, $R_n(\hat{m})$ is random for any data-dependent model selection procedure $\hat{m}$, so that the corresponding asymptotic optimality approach (shortly, asymptotic $r_n$-optimality) differs from the asymptotic mean efficiency approach of Shibata (1983), which requires $EL_n(\hat{m})/r_n \to 1$ as $n \to \infty$. □
For the sake of simplicity, in the sequel we will always assume that both the pseudo-
true and the maximal model belong to the set of admitted model candidates, i.e.
m_0, m_1 \in M_n.

2.2 Consistency

In practice, model selection is usually done on the basis of some criterion, C_n(m), say. 
The resulting model selection procedure is then the minimizer of this criterion over the 
set M_n of competing models, that is

\[ \hat{m} \in \arg \min_{m \in M_n} C_n(m) . \]  

(2.4)

Most criteria are motivated by being a reasonably good estimator of some discrepancy 
D_n(m). However, it is e.g. well known that the consistency of C_n(m) as estimator 
of D_n(m) is not enough for getting a consistent model selection procedure. Sufficient 
conditions to achieve this property may be formulated as follows.

**Proposition 2.2** A model selection procedure \( \hat{m} \) defined by (2.4) is consistent, if 
P(\hat{m} \in M_n^0) \to 1 and, for some \( m^* \in M_n^* \),

\[ \sup_{m \in M_n^0 \setminus M_n^*} \left| \frac{C_n(m) - C_n(m^*)}{\rho_n(m, m^*)} \right| - 1 \xrightarrow{P} 0 \]  

(2.5)

as \( n \to \infty \), where \( \rho_n(m, m^*) \) is some function satisfying \( \rho_n(m, m^*) > 0 \) for all \( m \notin M_n^* \).

In many situations, \( \rho_n(m, m^*) = s_n[D_n(m) - D_n(m^*)] \) provides a reasonable specification with some positive sequence \( s_n \) (possibly depending on \( m \) and \( m^* \)). Then condition (2.5) requires that the discrepancy difference \( D_n(m) - D_n(m^*) \) (possibly multiplied by some factor) is consistently estimated by the corresponding difference of the criterion \( C_n(m) - C_n(m^*) \) (uniformly in \( m \in M_n^0 \setminus M_n^* \)).

We consider now the following general criterion which is well defined for known variance \( \sigma^2 \):

\[ C_n(m) = RSS(m) + h_n \frac{|m|}{n} \sigma^2 , \quad h_n \geq 0 . \]  

(2.6)

The case of an unknown error variance is treated in Section 3. In accordance with 
Foster & George (1994), minimizers of (2.6) may be called canonical model selection 
procedures.

First we apply Proposition 2.2 to give conditions under which the procedure \( \hat{m} \) 
minimizing (2.6) is \( M_0 \)-consistent, i.e. the probability that it selects a model with
minimal bias tends to one as the sample size approaches infinity (cp. Example 2). Note that none of the competing models is assumed to be adequate, so that \( m_0 \) denotes in general the pseudo-true model.

**Theorem 2.1** A model selection procedure in minimizing (2.6) is \( M_0 \)-consistent, if

\[
\frac{\max\{|M_n \setminus M_0|, p_n h_n, p_n\}}{n \delta_n^*} \to 0 \quad \text{as} \quad n \to \infty .
\]  

(2.7)

In the special case that the covariates are given in a decreasing order of importance such as in polynomial regression, it is quite natural to consider only the case of nested models, i.e. \( M_n = \{\{1\}, \{1, 2\}, \ldots, \{1, 2, \ldots, p_n\}\} =: M_n^N \). Then the condition in the above Theorem may be weakened.

**Corollary 2.1** Let \( M_n = M_n^N \) and suppose that, as \( n \to \infty \),

\[
|m_0| \max\{h_n, 1\} = o(n \delta_n^*)
\]

(2.8)

Then a model selection procedure defined by (2.4) and (2.6) is \( M_0 \)-consistent.

Roughly speaking, conditions (2.7) and (2.8) prevent underfitting and are only achievable if \( n \delta_n^* \to \infty \) as \( n \to \infty \). Recall that \( \delta_n^* \) describes a measure for identifying pseudo-inadequate models from the pseudo-adequate ones, compare (2.3). The required growth rate of \( n \delta_n^* \) depends in the general case on quantities like the complexity of the considered class of model candidates, the maximal model dimension \( p_n \) and the penalty \( h_n \) of the criterion (2.6). On the other hand, in the special case of a bounded number of nested model candidates, condition (2.8) is satisfied by any diverging sequence \( n \delta_n^* \) if \( h_n \) is also bounded. Clearly, if the penalty \( h_n \) is bounded, then conditions (2.7) and (2.8), respectively, reduce to

\[
\max\{|M_n \setminus M_0|, p_n\} = o(n \delta_n^*) \quad \text{and} \quad |m_0| = o(n \delta_n^*)
\]

(2.9)

(2.10)

both as \( n \to \infty \).

Sufficient conditions for the \( m_0 \)-consistency of a model selection procedure can be derived in a similar way (see Example 1). Here, Theorem 2.1 is used to ensure, as required by Proposition 2.2, that the probability of selecting a model from \( M_n^0 = M_0 \) tends to one.
Theorem 2.2 A model selection procedure \( m \) defined by (2.4) and (2.6) is \( m_0 \)-consistent, if in addition to (2.7) the following condition holds:

\[
\min\{|M_0|, p_n - |m_0|\}/h_n \to 0 \text{ as } n \to \infty .
\]  

(2.11)

Condition (2.11) remedies the overfitting problem and is obviously fulfilled if, for example, \( p_n \) is bounded and \( h_n \to \infty \) as \( n \to \infty \). In the case of nested model candidates, the latter requirement is already sufficient under some moment condition.

Corollary 2.2 Let \( M_n = M_n^N \) and suppose that \( E\varepsilon_1^N < \infty \). Then a model selection procedure defined by (2.4) and (2.6) is \( m_0 \)-consistent if it is \( M_0 \)-consistent and \( h_n \to \infty \) as \( n \to \infty \).

The literature on the consistency of model selection procedures focuses mainly on \( m_0 \)-consistency and sometimes on the related \( M_0 \)-consistency. Nevertheless, the next theorem presents sufficient conditions, which ensure that a \( m_0 \)-consistent procedure is also \( M_n^*(D_n, M_n^0) \)-consistent for some choices of \( D_n \) and \( M_n^0 \). The idea is to show that \( m_0 \) minimizes \( D_n(m) \) over \( m \in M_n^0 \) for sufficiently large \( n \) (almost surely).

Theorem 2.3 Any \( m_0 \)-consistent model selection procedure is \( M_n^*(L_n, M_0) \)-consistent. It is also \( M_n^* \)- and \( M_l^* \)-consistent, if condition (2.10) is satisfied.

We recall that condition (2.10) is an immediate consequence of (2.7) or (2.8) and therefore not very restrictive for \( m_0 \)-consistent procedures.

2.3 Asymptotic optimality

The literature on the asymptotic optimality of model selection procedures is mainly concerned with situations as in Examples 3 and 4. A first result follows directly from the theorems and corollaries in the previous subsection by applying Proposition 2.2.

Corollary 2.3 A model selection procedure defined by (2.4) and (2.6) is asymptotically \( l_n \)- and \( r_n \)-optimal if it fulfills the assumptions of Theorem 2.2 or Corollary 2.2.

Asymptotic optimality of a procedure is also achievable under other conditions than above. For proving corresponding results, the following fact is especially useful.
Proposition 2.3 .

(i) A model selection procedure \( \hat{m} \) defined by (2.4) is asymptotically optimal, if for some \( m^* \in M^* \)

\[
\sup_{m \in M_n} \left| s_n[C_n(m) - C_n(m^*)] - \frac{D_n(m) - d_n}{D_n(m) + d_n} \right| \xrightarrow{P} 0 \quad (2.12)
\]

as \( n \to \infty \), where \( s_n \) is some positive sequence possibly depending on \( m \) and \( m^* \).

(ii) Let \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \) and suppose that

\[
\sup_{m \in M_n} \left| \frac{Z_n(m)}{D_n(m)} - 1 \right| \xrightarrow{P} 0 , \text{ as } n \to \infty , \text{ where} \quad (2.13)
\]

\[
Z_n(m) = L_n(m) + \frac{2}{n} \varepsilon^T (I - P_m) \mu - \frac{2}{n} \sigma^2 P_m \varepsilon + h_n \frac{|m|}{n} \mu^2 . \quad (2.14)
\]

Then a model selection procedure \( \hat{m} \) defined by (2.4) and (2.6) is asymptotically optimal.

In case of a known error variance, we apply now this proposition to derive conditions under which a model selection procedure based on criterion (2.6) is asymptotically optimal in the sense of Examples 3 and 4. For this we will additionally assume the existence of higher moments for the error distribution:

\[
E \varepsilon_1^q < \infty \quad \text{for some natural number } q \quad (2.15)
\]

Theorem 2.4 Assume that besides (2.15) the following conditions are satisfied as \( n \to \infty \):

\[
\sum_{m \in M_n} [n R_n(m)]^{-q} \to 0 \quad (2.16)
\]

\[
\sup_{m \in M_n} \frac{|h_n - 2||m|}{n R_n(m)} \to 0 . \quad (2.17)
\]

Then a model selection procedure defined by (2.4) and (2.6) is asymptotically optimal with respect to both \( l_n \) and \( r_n \).

Condition (2.17) is obviously implied by

\[
\frac{|h_n - 2| p_n}{n r_n} \to 0 \quad \text{or} \quad |h_n - 2| \to 0 \quad \text{as } n \to \infty , \quad (2.18)
\]

since \( n R_n(m) \geq |m| \sigma^2 \) for all \( m \). Moreover, (2.16) follows if \( |M_n| = o(|r_n|) \), which is e.g. fulfilled if \( |m^*| \) or \( n \delta_n \) diverge sufficiently fast to infinity compared with \( |M_n| \), i.e.
\[ |M_n| = o(\max\{|m^*|, n\delta_n\}|^q) \], where \( m^* \) denotes a minimizer of \( R_n(m) \) over \( m \in M_n \). Note that, as commented by Li (1987), in nonparametric regression as well as in the inadequate parametric case, \( nr_n \) will typically be of order \( n^\eta \) for some \( \eta > 0 \), so that (2.16) is achievable if \( |M_n| \) is of polynomial order. In the case of nested model candidates, it is again possible to weaken the assumptions.

**Corollary 2.4** Let \( M_n = M_n^N \). Then a model selection procedure defined by (2.4) and (2.6) is asymptotically optimal with respect to both \( l_n \) and \( r_n \), if (2.15) with \( q = 2 \), (2.17) and
\[
    nr_n \to \infty \quad \text{as} \quad n \to \infty. \tag{2.19}
\]

For instance, from Shibata (1981) it is known that in the problem of selecting an appropriate order in polynomial regression, condition (2.19) will hold when the true regression function is not a polynomial of any finite order.

**Remark.** The results of this subsection suggest that criteria (2.6) with penalties \( h_n = 2 \), at least approximately or in some asymptotic sense, should be the first candidates for providing asymptotically optimal procedures. Therefore, in Subsection 3.3 emphasis will be on such criteria. Nevertheless, condition (2.17), or (2.18), could also be satisfied for criteria with other penalties \( h_n \) if the maximal model dimension \( p_n \) increases very slowly with \( n \). To illustrate this fact, let us consider the simple case of nested models with algebraically decaying biases, i.e. \( \Delta_n(m) = C|m|^{-\eta} \) for some constants \( C, \eta > 0 \), and finite \( \ell \)th moments of the error distribution. Then \( nr_n \) is of order \( n^{1/(1+\eta)} \), so that any procedure based on a criterion (2.6) will be asymptotically optimal, provided that \( \lim_{n \to \infty} h_n p_n n^{-1/(1+\eta)} = 0 \) holds. Hence, even diverging sequences \( h_n \) are not completely excluded.

To bring this discussion to a head, let us finally assume that \( p_n \) is bounded and that all models in \( M_n \) are inadequate. Then \( \delta_n = \delta > 0 \), and any sequence \( h_n \) with \( \lim_{n \to \infty} h_n/n = 0 \) satisfies (2.18) and provides thus an asymptotically optimal procedure. The reason is, of course, that in such case the estimation error is dominated by the model bias, so that the largest model \( m_1 \) becomes optimal, e.g. with respect to the risk, for sufficiently large sample sizes. Notice that this trivial case is one of the rather rare situations where part (ii) of Proposition 2.1 is applicable, with \( D_n(m) = R_n(m) \) and \( M_n^0 = M_n \), since condition (2.2) will typically be fulfilled. That is, the asymptotic optimality of a procedure yields its \( M^*_\eta \)-consistency and consequently, on account of \( M^*_\eta = \{m_1\} = \{m_0\} \) for sufficiently large \( n \), its \( m_0 \)-consistency. Naturally, this “result” might be obtained directly, and it is not in contrast to Subsection 2.4 where \( h_n \to \infty \)
will turn out to be a necessary condition for \( m_0 \)-consistency, provided that neither \( m_0 \) nor \( m_1 \setminus m_0 \) are empty.

\[ \square \]

2.4 Illustration in case of orthonormal regressors  

1. (Assumptions) We assume that the errors in (1.1) are normally distributed, and that the basis functions \( g_i \) (\( i = 1, \ldots, p_n \)) are orthonormal in the summation sense, i.e.

\[
\frac{1}{n} \sum_{k=1}^{n} g_i(x_k)g_j(x_k) = \delta_{ij},
\]

where \( \delta_{ij} \) denotes the Kronecker symbol. Then we have, for any \( m \in M_n, G_m^T G_m = n I \), and the least squares estimates of the parameters,

\[
\hat{\beta}_i^{(m)} = \frac{1}{n} \sum_{k=1}^{n} y_k g_i(x_k) =: \beta_i \quad (i \in m),
\]

do not depend on the underlying model \( m \). Similarly, the components of the projection parameter \( \beta_i^f \) defined in (2.1) are given by \( \beta_i^f = n^{-1} \sum_{k=1}^{n} f(x_k)g_i(x_k) \) (\( i \in m_1 \)). The orthonormality assumption (2.20) provides

\[
\Delta_n(m) - \delta_n = \sum_{i \in m_1 \setminus m} (\beta_i^f)^2.
\]

Moreover, it is easily seen that the criterion (2.6) is minimized (over all \( m \subseteq m_1 \)) by

\[
\hat{m} = \{ i \in m_1 \mid \hat{\tau}_i^2 > h_n \},
\]

where \( \hat{\tau}_i = \sqrt{n} \hat{\beta}_i / \sigma \), cf. Droge (1993). Our assumptions ensure that these coefficients are independent distributed with

\[
\hat{\tau}_i \sim N(\tau_i, 1), \quad \tau_i = \sqrt{n} \beta_i / \sigma.
\]

2. (Necessary conditions for \( m_0 \)-consistency) The sufficient conditions for the \( m_0 \)-consistency of a canonical model selection procedure in Subsection 2.2 require at least that both \( h_n \) and \( n \delta_n^* \) tend to infinity as \( n \to \infty \) (and the divergence of \( h_n \) is slower than that of \( n \delta_n^* \)). In the considered special case we will illustrate that these conditions are necessary. Recalling the definition of \( m_0 \), our assumptions lead to

\[
p_n^0(i) := P(\hat{\tau}_i^2 > h_n) = 1 - \Phi(\sqrt{h_n - \tau_i}) + \Phi(-\sqrt{h_n - \tau_i}) \quad \text{for } i \in m_0,
\]

\[
p_n^*(i) := P(\hat{\tau}_i^2 \leq h_n) = 1 - 2\Phi(-\sqrt{h_n}) \quad \text{for } i \in m_1 \setminus m_0,
\]

14
where $\Phi$ denotes the distribution function of the standard normal law. Therefore,

$$p(n) := P(\hat{m} = m_0) = \prod_{i \in m_0} p^0(i) \prod_{i \in m_1 \setminus m_0} p^*(i)$$

must converge to one for any $m_0$-consistent procedure $\hat{m}$. Supposing additionally that neither $m_0$ nor $m_1 \setminus m_0$ are empty (for sufficiently large $n$), we obtain

$$p(n) \leq \min \{p^0(i^0), p^*(i^*)\} \quad \text{for any } i^0_n \in m_0, \quad i^*_n \in m_1 \setminus m_0.$$  \hspace{1cm} (2.23)

Taking $i^0_n \in \arg \min_{i \in m_0} \tau_i^2$ we get $n\delta_n^* = \sigma^2 \tau_{i^0_n}^2$, compare (2.21), so that the right hand side of inequality (2.23) tends to one if and only if

$$h_n \to \infty \quad \text{and} \quad \limsup_{n \to \infty} h_n/(n\delta_n^*) < 1,$$ \hspace{1cm} (2.24)

which implies in particular $n\delta_n^* \to \infty$ as $n \to \infty$.

3. (Sufficient conditions for $m_0$-consistency) We observe first that the necessary conditions (2.24) are also sufficient for the $m_0$-consistency of the procedure $\hat{m}$ if the maximal model dimension $p_n$ is bounded.

For the general case let $F_k(\cdot;\alpha)$ and $M_k(t)$, respectively, denote the distribution function and the moment generating function of a noncentral $\chi^2$-distributed random variable with $k$ degrees of freedom and noncentrality parameter $\alpha \geq 0$. Using the fact that $F_1(x;\alpha)$ is monotonically decreasing in $\alpha$, we obtain, for $i \in m_0$,

$$p^0(i) = 1 - F_1(h_n;\tau_i^2) \geq 1 - F_1(h_n; n\delta_n^*).$$

Together with the Bernoulli inequality, this leads to

$$p(n) \geq [1 - |m_0| F_1(h_n; n\delta_n^*)] \left[ 1 - 2 (p_n - |m_0|) \Phi(-\sqrt{h_n}) \right] \geq \left[ 1 - \frac{|m_0|}{\sqrt{3}} \exp(h_n - n\delta_n^*/2) \right] \left[ 1 - \frac{\sqrt{2} (p_n - |m_0|)}{\sqrt{\pi h_n}} \exp(-h_n/2) \right].$$

To establish the last line we have used the Markov inequality providing

$$F_k(h_n; n\delta_n^*) \leq \exp(h_n) M_k(-1) = 3^{-k/2} \exp(h_n) \exp(-n\delta_n^*/3)$$ \hspace{1cm} (2.25)

as well as the elementary relation $\Phi(-x) \leq \varphi(x)/x$, where $\varphi$ denotes the density function of the standard normal distribution. Consequently, the procedure $\hat{m}$ is $m_0$-consistent, that it, $p(n) \to 1$ as $n \to \infty$, if $h_n \to \infty$ and

$$\max \left\{ \limsup_{n \to \infty} \frac{\ln(p_n - |m_0|)}{h_n}, \limsup_{n \to \infty} \frac{h_n + \ln(|m_0|)}{n\delta_n^*} \right\} < \frac{1}{2}.$$
Since $m_0$ is unknown, this provides $2\ln(p_n)$ as lower bound for the dimensionality penalty $h_n$. We remark that $h_n = 2\ln(p_n)$ is just the proposal of Foster and George (1994) in the orthogonal regressor case, which was derived by a different approach using the so-called risk inflation criterion.

4. (Case of an adequate model) We suppose now that $f$ may be expressed as

$$f = \sum_{j=1}^{k_0} \beta_{ij}^0 g_j,$$

where the nonvanishing coefficients $\beta_{ij}^0 (j = 1, \ldots, k_0)$ do not depend on $n$, and that $p_n \geq i_{k_0}$ for sufficiently large $n$ ($n \geq n_0$, say). Then we obtain

$$\mu = G\beta f, \quad \delta_n = 0 \quad \text{and} \quad \delta_n^* \geq \min \{|\beta_{ij}^0|^2 : j = 1, \ldots, k_0\},$$

providing, for example, $n\delta_n^* \to \infty$ as $n \to \infty$. The results of the last paragraph show that the procedure $\hat{m}$ defined by (2.22) is $m_0$-consistent if, as $n \to \infty$, $h_n \to \infty$, $h_n/n \to 0$ and $\ln(p_n)/h_n \to 0$. Corollary 2.4 implies immediately that $\hat{m}$ is asymptotically $l_n$- and $r_n$-optimal under the same assumptions. Note that $p_n$ may be bounded or not, and we have here $h_n \to \infty$ as well as, for sufficiently large $n$, $nr_n = k_0\sigma^2 < \infty$. The last fact is shown in the Appendix, where we also establish the asymptotic mean efficiency (cf. Example 4) of the procedure. On the other hand we will see there that model selection procedures with any bounded penalty sequence $h_n$ can also be asymptotically mean efficient under different assumptions, e.g. if in the inadequate case $\lim_{n \to \infty} p_n/(nr_n) = 0$ holds.

3 Application to some criteria

3.1 A general representation of model selection criteria

In practice, the variance $\sigma^2$ will be unknown. Therefore we consider the following model selection criterion instead of (2.6):

$$\hat{C}_n(m) = RSS(m) + \hat{h}_n(m) \frac{|m|}{n}\sigma^2,$$

where the (nonnegative) stochastic penalty, $\hat{h}_n(m)$, may depend on $m$. Any minimizer of the criterion $\hat{C}_n(m)$ over $m \in M_n$ will be called minimum-$\hat{C}_n$-procedure. Representations (3.1) exist for many of the popular model selection criteria, typically with

$$\hat{h}_n(m) = h_n(m)\hat{\sigma}_n^2(m)/\sigma^2,$$
where $h_n(m) \geq 0$ is a nonrandom penalty and $\hat{\sigma}^2_n(m)$ is some “variance estimator” under model $m$. Table 3.1 contains the corresponding expressions for $h_n(m)$ and $\hat{\sigma}^2_n(m)$ for all criteria introduced in Section 1. There the following notations are used:

$$\hat{\sigma}^2(m) = \frac{n}{n - |m|} RSS(m)$$

$$\Gamma(m) = \text{diag} \gamma_1(m), \ldots, \gamma_n(m) , \gamma_i(m) = p_{ii}(m) \{2 - p_{ii}(m)\}^{-1} \{1 - p_{ii}(m)\}^{-2}$$

$$\Lambda(m) = \text{diag} \lambda_1(m), \ldots, \lambda_n(m) , \lambda_i(m) = p_{ii}(m) \{2 + p_{ii}(m)\}^{-1}$$

Table 3.1: Representation of model selection criteria by (3.1), (3.2)

<table>
<thead>
<tr>
<th>Criterion</th>
<th>$h_n(m)$</th>
<th>$\hat{\sigma}^2_n(m)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_p$</td>
<td>2</td>
<td>$\hat{\sigma}^2(m_1)$</td>
</tr>
<tr>
<td>FPE</td>
<td>2</td>
<td>$\hat{\sigma}^2(m)$</td>
</tr>
<tr>
<td>SH</td>
<td>$2(1 - n^{-1}</td>
<td>m</td>
</tr>
<tr>
<td>GCV</td>
<td>$2 - (n -</td>
<td>m</td>
</tr>
<tr>
<td>GFCV</td>
<td>$2 - n^{-1}</td>
<td>m</td>
</tr>
<tr>
<td>$\exp(GIC)$</td>
<td>$a_n q_n(m)(1 - n^{-1}</td>
<td>m</td>
</tr>
<tr>
<td>CV</td>
<td>$1 +</td>
<td>m</td>
</tr>
<tr>
<td>FCV</td>
<td>$2 -</td>
<td>m</td>
</tr>
</tbody>
</table>

Appropriate specifications of $a_n$ in the definition of GIC give the corresponding representations for AIC, BIC and $\phi$, cf. Section 1. Note that $q_n(m) \rightarrow 1$ if $a_n n^{-1}|m| \rightarrow 0$. Moreover, all estimators $\hat{\sigma}^2_n(m)$ in Table 3.1 are unbiased for $\sigma^2$, provided that the model $m$ is adequate (correct), i.e. $\Delta_n(m) = 0$. This is an immediate consequence of the following lemma which summarizes some properties of the variance estimators.

**Lemma 3.1** Let $\Omega_m = \text{diag} \omega_1(m), \ldots, \omega_n(m)$ with $\omega_i(m) \geq 0$ for $i = 1, \ldots, n$, $\omega(m) = \max_{i=1}^n \omega_i(m)$, $T_m = (I - P_m)\Omega_m(I - P_m)$, $t_m = \text{tr}(T_m) = \text{tr}[\Omega_m(I - P_m)]$ and $\hat{\sigma}^2_n(m) = t_m^{-1}\|y\|_T^2$. Then we have, for any $m \in M_n$,

(i) $E\hat{\sigma}^2_n(m) = \sigma^2 + t_m^{-1}\|\mu\|_T^2 =: e_n(m)$ and,

(ii) assuming additionally (2.15), $E|\hat{\sigma}^2_n(m) - e_n(m)|^q = O([t_m^{-1}e_n(m)\omega(m)]^q)$ .

17
The result for $\hat{\sigma}^2(m)$ follows by setting $\Omega_m = I$, so that $\omega(m) = 1$, $t_m = n - |m|$ and $e_n(m) = \sigma^2 + \frac{n}{n-|m|} \Delta_n(m)$. The “variance estimators” occurring in the CV and FCV criteria are covered by $\Omega_m = \Gamma(m)$ and $\Omega_m = \Lambda(m)$, respectively.

It may intuitively be expected that a minimizer of (3.1) will share the properties of a minimizer of (2.6) if the stochastic penalty $\hat{h}_n(m)$ behaves similarly to the penalty $h_n$ in (2.6). The next two subsections make this precise in the context of both consistency and asymptotic optimality.

### 3.2 Consistency of some procedures

Here, we restrict our investigations to $M_0$- and $m_0$-consistency. For each of the criteria in Table 3.1 we will first present conditions, which ensure the $M_0$-consistency of the associated model selection procedure. We denote by

$$h_n := \sup_{m \in M_n} h_n(m) \quad \text{and} \quad h^0_n := \sup_{m \subseteq m_0} h_n(m)$$

(3.3)

the maximum nonrandom penalty terms over $m \in M_n$ and $m \subseteq m_0$, respectively, and start with a general result on procedures minimizing (3.1), where either $\hat{\sigma}^2(m)$ or $\hat{\sigma}^2(m_1)$ serves as variance estimate.

**Theorem 3.1** Let $\hat{m}$ be a minimizer of (3.1), where $\hat{h}_n(m)$ is given by (3.2).

(i) Assume that (2.7) holds for $h_n := \hat{h}_n E \hat{\sigma}^2(m_1)/\sigma^2$. Then $\hat{m}$ is $M_0$-consistent, if

(a) $\hat{\sigma}^2(m) = \hat{\sigma}^2(m_1)$ or,

(b) $\hat{\sigma}^2(m) = \sigma^2(m)$, and $h_n p_n/(n - p_n) \to 0$ as $n \to \infty$.

(ii) Let $M_n = M_n^N$ and assume that (2.8) holds for $h_n := h^0_n E \hat{\sigma}^2(m_0)/\sigma^2$. Then $\hat{m}$ is $M_0$-consistent, if either (a) or (b), but with $|m_0|$, $\hat{h}^0_n$ instead of $p_n, \hat{h}_n$, is satisfied.

In both cases (a) and (b) we have $\hat{\sigma}^2(m_1) = \hat{\sigma}^2(m_1)$, so that condition (2.7) implies that the bias of $\hat{\sigma}^2(m_1)$, $E \hat{\sigma}^2(m_1) - \sigma^2 = n \delta_n/(n - p_n)$, is asymptotically negligible compared with $n \delta_n/(\hat{h}_n p_n)$, that is,

$$\frac{\hat{h}_n p_n \delta_n}{(n - p_n) \delta_n^*} \to 0 \quad \text{as} \quad n \to \infty \quad .$$

(3.4)

This property follows also from the condition in (b), if

$$\delta_n = O(\delta_n^*) \quad \text{as} \quad n \to \infty \quad .$$

(3.5)
Usually, $h_n \geq 0$ will be bounded away from zero. Then the condition in (b) requires that the largest model dimension increases slower than the sample size,

$$p_n = o(n) \quad \text{as} \quad n \to \infty ,$$

so that the condition in (b) could be rewritten as

$$\bar{h}_np_n = o(n) \quad \text{as} \quad n \to \infty .$$

Theorem 3.1 can be applied to all criteria in Table 3.1, except for CV and FCV, and to the so-called generalized $C_p$-criterion, see, for example, Atkinson (1980) and Zheng and Loh (1995). This criterion may be expressed in terms of the general criterion (3.1) by taking $\hat{h}_n(m) = \hat{h}_n = \kappa_n \hat{\sigma}^2(m_1)/\sigma^2$, where $\kappa_n$ is some nonrandom positive sequence, depending possibly on $n$:

$$GC_p(m) = RSS(m) + \kappa_n \frac{|m|}{n} \hat{\sigma}^2(m_1).$$

The usual $C_p$-criterion (1.7) corresponds to the particular choice $\kappa_n \equiv 2$.

**Corollary 3.1** (i) Suppose that (2.9) and (3.4) with $\bar{h}_n := 2$ hold. Then the minimum-$C_p$-procedure is $M_0$-consistent. The same property is shared by the procedures based on the criteria AIC, FPE, SH, GCV and GFCV if additionally (3.6) is assumed.

(ii) Minimizing the criteria GIC and $GC_p$ provides also $M_0$-consistent procedures, if conditions (2.7), (3.4) and, in case of GIC, (3.7) are satisfied, where both $h_n$ and $\bar{h}_n$ have always to be replaced by $a_n$ (for GIC) or $\bar{a}_n$ (for $GC_p$).

(iii) In the special case of nested model candidates, $M_n = M_n^N$, the results remain true if condition (2.9) and (2.7) are replaced by (2.10) and (2.8), respectively, and if $|m_0|$ is used instead of $p_n$ in (3.6) (or in (3.7) for GIC) and (3.4) (but for $C_p$ and $GC_p$ only in the nominator of (3.4)).

We close our $M_0$-consistency considerations by dealing with the procedures based on the cross-validation and the full cross-validation criteria. Let

$$c_n(m) := \sup_{i=1}^n P_{n_i}(m)$$

denote the maximum diagonal element of the hat matrix (projection) $P_m$.

**Theorem 3.2** Suppose that (2.9) and, as $n \to \infty$,

$$c_n := \sup_{m \in M_n} c_n(m) \to 0 \quad \text{and}$$

$$c_n(\sigma^2 + \delta_n)/\delta_n^* \to 0$$

19
are satisfied. Then the model selection procedures minimizing $CV(m)$ and $FCV(m)$ over $m \in M_n$ are $M_0$-consistent. In the special case of nested model candidates, the result remains true if in (3.10) $m \in M_n$ is replaced by $m \subseteq m_0$, and if condition (2.10) is assumed instead of (2.9).

Notice that under (3.5) the condition (3.11) follows from (3.10), which in turn is a consequence of (3.6) if extremely unbalanced designs are excluded by assuming, for example,

$$\exists K > 0 \forall n \forall m \in M_n \quad c_n(m) \leq K \frac{|m|}{n}.$$ (3.12)

Our sufficient conditions for the $m_0$-consistency of canonical model selection procedures in Subsection 2.2 require that the nonrandom penalty $h_n$ diverges to infinity as $n \to \infty$, compare (2.11) and Corollary 2.2. In case of procedures minimizing the criterion (3.1) we would expect that the related substitute $\tilde{h}_n$ defined in Theorem 3.1 shares this property. Among the criteria considered above, GIC and $GC_p$ appear therefore as the most appropriate candidates to achieve $m_0$-consistency.

To study the behaviour of the minimum-GIC-procedure, we observe first that

$$1 \leq q_n(m) \leq q_n(m_1) =: q_n \text{ for all } m \in M_n,$$

see Table 3.1. This gives $\tilde{h}_n = h_n(m_1) = a_n q_n(1 - p_n/n)$ and hence, on account of Lemma 3.1,

$$h_n = E\tilde{h}_n(m_1) = a_n q_n E[RSS(m_1)]/\sigma^2 = a_n q_n (\delta_n/\sigma^2 + 1 - p_n/n),$$ (3.13)

compare (3.3) and Theorem 3.1. Typically, the minimal model bias $\delta_n$ will be bounded, i.e.

$$\delta_n = O(1) \quad \text{as} \quad n \to \infty,$$ (3.14)

or it will even vanish asymptotically. Then the “minimal” requirement $h_n \to \infty$ leads to $a_n q_n \to \infty$ and thus to $a_n \to \infty$, since the sequence $\{q_n\}$ is bounded if $\{a_n\}$ is bounded. Together with the sufficient conditions for the $M_0$-consistency in Corollary 3.1, e.g. (3.7) implying (3.6), this establishes the weak asymptotic equivalence of $h_n$ and $a_n$.

**Theorem 3.3** Suppose that conditions (3.14), (2.7), (2.11) and (3.7) are satisfied, with $a_n$ instead of both $h_n$ and $\tilde{h}_n$. Then the model selection procedure defined as minimizer of the GIC criterion is $m_0$-consistent.

In the special case of nested model candidates, the assumptions can be weakened in the same way as for the canonical model selection criterion.
Corollary 3.2 Let $M_n = M^N_n$, and suppose that $E\varepsilon_1^i < \infty$, (3.14) and (3.7), with $a_n$ instead of $\tilde{h}_n$, hold. Then the minimum-GIC-procedure is $m_0$-consistent if, as $n \to \infty$, $a_n \to \infty$ and

$$\frac{a_n|m_0|}{n\delta_n} \to 0.$$ (3.15)

We conclude this subsection by investigating the asymptotic properties of the minimum-GC$_p$-procedure. For GC$_p$ we have $\tilde{h}_n = \kappa_n$ and

$$h_n = E\tilde{h}_n(m_1) = \kappa_n E\hat{\sigma}^2(m_1)/\sigma^2 = \kappa_n \left(1 + \frac{n\delta_n}{(n - p_n)\sigma^2}\right),$$ (3.16)

compare (3.3) and Theorem 3.1.

Theorem 3.4 Suppose that

$$(n - p_n)E\hat{\sigma}^2(m_1 = (n - p_n)\sigma^2 + n\delta_n \to \infty \text{ as } n \to \infty,$$ (3.17)

and one of the following two sets of conditions, with $h_n$ defined by (3.16), are satisfied:

(i) $\sigma^2 = \varepsilon_1^i$, (2.7), (2.11), (2.15) with $q = 1$, or

(ii) $M_n = M^N_n$, (2.8), (2.15) with $q = 2$, and $h_n \to \infty$ as $n \to \infty$.

Then the model selection procedure minimizing (3.8) over $m \in M_n$ is $m_0$-consistent.

Notice that condition (3.17) is fulfilled if $(n - p_n)$ diverges to infinity as $n \to \infty$, which is fairly minimal in proving any asymptotic theory. Nevertheless, (3.17) could be achieved even in cases where $(n - p_n)$ is bounded, provided that $n\delta_n \to \infty$ as $n \to \infty$.

3.3 Asymptotic optimality of some procedures

Similarly to Corollary 2.3, the results of the previous subsection provide the following.

Corollary 3.3 The minimum-GIC-procedure is asymptotically $l_n$- and $r_n$-optimal if the conditions of Theorem 3.3 or Corollary 3.2 are satisfied. Moreover, the minimum-GC$_p$-procedure shares this optimality property under the assumptions of Theorem 3.4.

This result will typically be applicable in situations where $nr_n$ is bounded.

Next we are concerned with optimality conditions for the minimizer of the general criterion $\hat{C}_n$ when $nr_n$ diverges to infinity.
Proposition 3.1 Assume that (2.15), (2.16) and the following condition are fulfilled:

$$\sup_{m \in M_n} \frac{\hat{h}_n(m) - 2\|m\|\sigma^2}{nR_n(m)} \xrightarrow{P} 0 \text{ as } n \to \infty.$$  \hspace{1cm} (3.18)

Then a model selection procedure \( \hat{m} \) minimizing (3.1) over \( m \in M_n \) is asymptotically \( r_n \)- and \( I_n \)-optimal.

Recalling \( |m|\sigma^2 \leq nR_n(m) \), condition (3.18) is obviously fulfilled if \( \hat{h}_n(m) \) converges in probability to 2, uniformly in \( M_n \). For example, this holds for the \( C_p \)-criterion if \( \hat{\sigma}^2(m_1) \) is a consistent estimate of \( \sigma^2 \). However, this consistency of \( \hat{\sigma}^2(m_1) \) is not a necessary condition for the optimality of the minimum-\( C_p \)-procedure as stated in the next theorem.

Theorem 3.5 The minimum-\( C_p \)-procedure is asymptotically \( r_n \)- and \( I_n \)-optimal if (2.15), (2.16) and one of the following conditions are satisfied: (3.6) or

$$\limsup_{n \to \infty} p_n/n < 1 \text{ and } \delta_n = o(r_n) \text{ or } \frac{n - p_n}{n - p_n} \to \infty \text{ as } n \to \infty. \hspace{1cm} (3.19)$$

Concerning (3.19) we note that \( \delta_n = o(r_n) \) is implied by \( \delta_n = o(|m^*|n^{-1}) \), where again \( m^* \in \arg\min_{m \in M_n} R_n(m) \), since \( r_n \geq |m^*|n^{-1}\sigma^2 \). Condition (3.19) is fulfilled if, for example, \( p_n = \gamma n \) for some \( \gamma \in (0, 1) \) and if the model biases are algebraically decaying as in the Remark of Subsection 2.3. Condition (3.20) provides the consistency of \( \hat{\sigma}^2(m_1) \) and requires that \( \delta_n \to 0 \) as \( n \to \infty \). This condition allows even \( \lim_{n \to \infty} p_n/n = 1 \), but only if the minimal model bias vanishes sufficiently fast. For instance, (3.20) is satisfied if \( p_n = n - \ln(n) \) and \( \delta_n = o(\ln(n)n^{-1}) \). On the other hand, assuming (3.6) the asymptotic optimality of the minimum-\( C_p \)-procedure may even be achieved in cases where \( \delta_n \) doesn’t converge to 0 (and hence \( \hat{\sigma}^2(m_1) \) is asymptotically biased).

We study now the behaviour of a class of model selection procedures which covers e.g. those based on the FPE and AIC criteria.

Theorem 3.6 Let \( \hat{m} \) be a minimizer of (3.1), where \( \hat{h}_n(m) \) is defined by (3.2) with \( \hat{\sigma}^2(m) = \hat{\sigma}^2(m) \). Assume that (2.15), (2.16) and the following conditions are satisfied:

$$\limsup_{n \to \infty} p_n/n < 1, \hspace{1cm} (3.21)$$

$$\limsup_{n \to \infty} \sup_{m \in M_n} h_n(m) < \infty. \hspace{1cm} (3.22)$$
\[
\lim_{n \to \infty} \sup_{m \in M_n} \frac{|h_n(m) - 2| m}{n R_n(m)} = \lim_{n \to \infty} \sup_{m \in M_n} \frac{|m| \Delta_n(m)}{(n - |m|) R_n(m)} = 0 .
\] (3.23)

Then \( m \) is asymptotically \( r_n \)- and \( l_n \)-optimal.

Because of \( n R_n(m) \geq |m| \sigma^2 \), the first condition in (3.23) is fulfilled if

\[
\sup_{m \in M_n} |h_n(m) - 2| \to 0 \text{ as } n \to \infty .
\] (3.24)

Another sufficient condition for this is \( p_n = o(n \delta_n) \), since \( |h_n(m) - 2| \) is uniformly bounded in \( M_n \) by (3.22) and \( |m|/(n R_n(m)) \leq p_n/(n \delta_n) \).

Recalling (1.4), we obtain

\[
\frac{|m| \Delta_n(m)}{(n - |m|) R_n(m)} \leq \min \left\{ \frac{|m|}{n - |m|}, \frac{n}{n - |m|} \Delta_n(m) \right\} .
\]

Hence, the second condition in (3.23) follows from (3.6) or if there exist a divergent sequence \( \{k_n\} \) with

\[
k_n \leq p_n , \ k_n = o(n) \text{ and } \sup_{|m| \geq k_n} \frac{n}{n - |m|} \Delta_n(m) \to 0 \text{ as } n \to \infty .
\] (3.25)

Notice that the last condition in (3.25) requires the asymptotic unbiasedness of \( \hat{\sigma}^2(m) \) uniformly in \( \{m \in M_n : |m| \geq k_n\} \).

Clearly, (3.24) holds always for the FPE criterion since then \( h_n(m) \equiv 2 \). Moreover, condition (3.24) may be verified under (3.6) for the criteria AIC, GCV, GFCV and SH, too. It cannot be expected to weaken this assumption by \( p_n = o(n \delta_n) \), since condition (3.14) will usually hold. Finally, (3.21) and (3.22) are implied by (3.6) and (3.24), respectively, so that we arrive at the following.

**Corollary 3.4** The model selection procedures defined as minimizers of the FPE, AIC, GCV, GFCV and SH criteria, respectively, are asymptotically \( r_n \)- and \( l_n \)-optimal if (2.15), (2.16) and (3.6) are satisfied. For the FPE criterion, the result remains true if the assumption (3.6) is replaced by (3.21) and (3.25).

The next theorem presents sufficient conditions for the asymptotic optimality of the model selection procedures based on cross-validation and full cross-validation.

**Theorem 3.7** Both the minimum-CV- and the minimum-FCV-procedures are asymptotically \( r_n \)- and \( l_n \)-optimal if conditions (2.15), (2.16) and (3.10) are satisfied.

Finally, we remark that all results of this subsection remain true in the special case of nested model candidates, if condition (2.15) is assumed for \( q = 2 \) and if (2.16) is replaced by (2.19). This is an immediate consequence of Corollary 2.4 and its proof.

23
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, which are closey related to our paper.

Nishii (1984) considered the problem of selecting an appropriate submodel of some given linear model, say $m_1 = \{1, \ldots, p\}$ 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 $\lim_{n \to \infty} n^{-1} G^T G$ exists and is also positive definite. Reformulating Nishii’s results in terms of our notions, he showed that under (N) and the assumption of normally distributed errors, procedures based on criteria CV, $C_p$, FPE and AIC are $M_\sigma$-consistent but not $m_\sigma$-consistent, that is, the selected models apt to overfit. In contrast, the $m_0$-consistency was proved for the criterion GIC under the additional assumption

$$a_n \to \infty \quad \text{and} \quad a_n = o(n) \quad \text{as} \quad n \to \infty.$$  \hfill (4.1)

Note that for the result on CV, $\lim_{n \to \infty} c_n(m_1) = 0$ is additionally required, which is in this case equivalent to our condition (3.10). Moreover, Nishii’s assumptions imply that neither $M_0$ nor $m_0$ depend on $n$ as well as that $\delta_n = 0$ and $\lim \inf_{n \to \infty} \delta_n^* > 0$. Thus, our conditions (2.9), (3.4), (3.6) and (3.14) are always satisfied, whereas (3.7), (2.7) and (2.11) follow from (4.1). This shows that Nishii’s results may be seen as special cases of our Theorems 3.2, 3.3 and Corollary 3.1.

The above results have been generalized by Müller (1993) to the case of nonnormal errors and inadequate linear models, defining an asymptotically true (instead of a true) model by $m_a = \{i \in m_1 \mid \lim \inf_{n \to \infty} |\beta_i| > 0\}$ and assuming some additional conditions on the design and the unknown regression function, including e.g. Nishii’s conditions on $G^T G$ as well as $||\mu||^2 = O(n)$. 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 $m_a$ is still fixed. Furthermore, assuming certain conditions to ensure that $m_a$ minimizes $L_n(m)$ for sufficiently large $n$, which are fulfilled e.g. under (N), $m_a$-consistent procedures turn out to be also asymptotically $l_n(r_n)$ optimal, which may be compared with our Corollary 3.3. Note that, in contrast to Müller (1993), we have tried to avoid imposing assumptions
on quantities like the asymptotic design (or information) matrix and the asymptotic projection parameters. Instead, our conditions are formulated in terms of quantities which are defined for each given sample size \(n\), such as the projection parameter \(\beta^T\), the pseudo-true model \(m_0\), the minimal model bias \(\delta_n\), and \(\delta_n^*\) defined by (2.3). Moreover, the separation between conditions on the design and the parameters is not always necessary, since one is often interested in estimating the regression function itself. We do also not assume that \(\|\mu\|^2/n\) is asymptotically bounded. Generally, our results appear more general than those of Müller (1993), who focused mainly on consistency.

In the situation of Nishii (1984) but with errors as in (1.1), 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-

\[\lim \inf_{n \to \infty} \delta_n^* > 0, \quad G^T G = O(n), \quad (G^T G)^{-1} = O(n^{-1}), \text{ and (3.10)}.\]

Zhang (1993) dealt with multifold CV in the same context. Under some assumptions including \(d/n \to \delta > 0\) as \(n \to \infty\) and (3.10), he established that the CV(\(d\)) criterion is asymptotically equivalent to the generalized \(C_p\)-criterion (3.8), where \(\kappa_n\) is replaced by \(\alpha = (2 - \delta)/(1 - \delta)\). Obviously, it holds \(\alpha > 2\) if \(\delta > 0\), whereas the criterion with \(\alpha = 2\) 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 \(d/n \to 1\) 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 \(\delta\). When \(\delta \to 0\), the CV(\(d\)) criterion becomes equivalent to the CV criterion.

Model selection procedures based on minimizing the criterion (3.8) 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 \(M_n^N\). The errors in (1.1) were assumed to be sub-Gaussian and, moreover, the maximal model dimension \(p_n\) was allowed to depend on \(n\), satisfying (3.21) (which implies condition (3.17)), whereas the true model did again not depend on the sample size. The authors considered some positive nondecreasing function, \(h_n(|m|)\), of \(|m|\) instead of \(\kappa_n|m|\) as the penalty term for the model complexity in (3.8). They showed
how this term has to be chosen 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, and the other authors mentioned above). In particular they allowed that, as in our paper, the minimal bias of an inadequate model, which is $\delta^*_n$ due to $\delta_n = 0$, may tend to zero, but not faster than $\rho_n \ln(n)/n$ for some sequence $\rho_n \to \infty$. The growth restrictions on $h_n$ reveal the interplay of $h_n$, $p_n$ and $\delta^*_n$, 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 $n\delta^*_n$ since, roughly speaking, $h_n$ has to increase faster than $p_n$ but slower than $n\delta^*_n$. Generally, if $p_n \to \infty$ as $n \to \infty$, then the penalty $h_n(|m|)$ is required to grow faster than when $p_n$ is bounded. The results of Zheng and Loh (1995) can be compared with part (ii) of Theorem 3.4. They use a somewhat more general penalty function $h_n(|m|)$, but a more specific restriction on the growth of $n\delta^*_n$. More importantly, our assumptions are weaker than theirs: Our result doesn’t require $\lim_{n \to \infty} p_n/\kappa_n = 0$, as it follows from condition A3 of Zheng and Loh, and we assume only the existence of finite 8th moments for the error distribution instead of sub-Gaussian errors, which have finite moments of all orders. We remark finally that Zheng and Loh (1995) have also established an almost sure version of the consistency result under some strengthened conditions.

It should be pointed out that all consistency results on which we have commented until now depend heavily on the assumed existence of a fixed finite-dimensional true (or asymptotically true) model. In contrast, we allow that the pseudo-true model depends on the sample size, which appears reasonable if the maximal model dimension may tend to infinity. Nevertheless, potential applications of our consistency results cover mainly cases where the dimension of the (pseudo-)true model is bounded. If, for example, the regression function can be expressed as an infinite series, $f = \sum_{j=1}^{\infty} \beta_j g_j$, with infinitely many nonvanishing coefficients, then zero-coefficients are often estimated with biases on the basis of any finite sample. These biases correspond to nonvanishing projection parameters and are in general small, implying that $n\delta^*_n$ will not tend to infinity. The consideration of an asymptotically true model is then no way out, since its definition may be difficult when the coefficients decrease with the related “frequency” $j$.

We have seen that the story is quite different when the dimension of the true model increases with the sample size or is infinite. Then it seems appropriate to use the asymptotic optimality approach instead of consistency considerations. Generally, in such situations procedures with comparatively small penalties for the model complexity
become preferable, which is in contrast to cases where the $m_0$-consistency approach works.

The first result on asymptotic optimality of linear model selection procedures is due to Shibata (1981), who established under certain conditions that procedures based on criteria such as SH, AIC, FPE and $C_p$ are asymptotically optimal, whereas those with larger penalties like BIC and $\phi$ are not. A procedure $\hat{m}$ was called asymptotically optimal, if $L_n(\hat{m})/r_n$ converges in probability to one. This concept is closely related to the asymptotic $L_n$-optimality, since the imposed assumptions usually imply that $L_n(m)/R_n(m)$ converges in probability to one, uniformly in $m \in M_n$, compare the proof of Theorem 2.4. Shibata assumed normally distributed errors in (1.1) as well as an infinite series expansion for the regression function, $f(x) = \sum_{j=1}^{\infty} \beta_j g_j(x)$, where both sequences $(\beta_1, \beta_2, \ldots)$ and $(g_1(x), g_2(x), \ldots)$ are square-summable. Moreover, the main conditions to achieve the asymptotic optimality of the procedures were (3.6) and

$$
\lim_{n \to \infty} \sum_{m \in M_n} \delta^n R_n(m) = 0 \quad \text{for any } \delta \in (0, 1). \quad (4.2)
$$

Notice that (4.2) implies condition (2.19), which is a minimal prerequisite for getting asymptotic $l_n$-optimality in Subsections 2.3 and 3.3. In a later paper, Shibata (1983) showed that his results remain true if the asymptotic optimality approach is replaced by the asymptotic mean efficiency approach as introduced in Example 4.

Li (1987) proved the asymptotic optimality of the procedures based on the criteria $C_p$, CV and GCV. The results of Droge (1999) on FCV as well as those in Subsections 2.3 and 3.3 are in the same spirit. More precisely, Li (1987) showed that, in our notation, the canonical model selection procedure with $h_n \equiv 2$ in (2.6) (called $C_p$) is asymptotically $l_n$-optimal under conditions (2.15) and (2.16). In the special case of nested model candidates, $M_n = M_n^N$, he demonstrated that it suffices to assume (2.15), with $q = 2$, and (2.19). These results are just stated in Theorem 2.4 and Corollary 2.4, with the only exception that we additionally require condition (2.17) to deal with general canonical model selection procedures; recall that (2.17) is automatically satisfied when $h_n = 2$. Thus, our proofs coincide essentially with those of Li (1987). In case of an unknown error variance $\sigma^2$, Li proposed to replace $\sigma^2$ by any consistent estimate. This keeps the optimality of the procedure, and is one of the possible versions provided by Theorem 3.5.

In Subsection 3.3 we have used the results on canonical model selection to derive the asymptotic $l_n$- and $r_n$-optimality of a variety of data-driven procedures. The proofs rely mainly on rewriting the underlying criterion as in (3.1) and applying Proposition 3.1.
In this way we have established the asymptotic optimality of the procedures based not only on $C_p$, but also on the criteria FPE, AIC, SH, GCV, CV, GFCV and FCV, where the last result was already published in Droge (1999). The ideas of Li (1987) in proving the asymptotic optimality of the procedures minimizing either CV or GCV are different. To illustrate the differences in the assumptions we restrict to the case of nested model candidates. Then Li used, instead of our condition (3.6), the following to achieve the asymptotic optimality of the minimum-GCV-procedure:

$$\inf_{m \in M_n^N} L_n(m) \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty$$

(4.3)

$$\exists C \forall a \geq 0 \sup_{x \in \mathbb{R}} P\{x - a \leq \varepsilon_1 \leq x + a\} \leq Ca .$$

(4.4)

Condition (4.3) requires in particular $\lim_{n \to \infty} \delta_n = 0$, whereas the condition (4.4) on the error distribution is satisfied if its density is bounded. To get the asymptotic optimality of the minimum-CV-procedure, Li (1987) replaced our assumption (3.10) by (4.3), (3.12) and $\lim_{n \to \infty} c_n < 1$, compare (3.10) for the definition of $c_n$ and note that $c_n < 1$ holds for any fixed $n$ (Droge, 1982, Lemma 2.2). Finally, 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: Proofs**

**Proof of Lemma 2.1.** We note first that, for any $m \in M_n$,

$$\Delta_n(m) = \Delta_n(m_1) + \frac{1}{n}\| (P - P_m) \mu \|^2 .$$

Consequently, we have $\Delta_n(m_1) = \min_{m \in M_n} \Delta_n(m) =: \delta_n$, and $\Delta_n(m) = \delta_n$ holds if and only if

$$(P - P_m) \mu = 0 .$$

(A1)

Using $P_m P = P_m$ and $P \mu = G \beta^T$, we obtain, with $G = (\bar{g}_1, \ldots, \bar{g}_{p_m})$,

$$(P - P_m) \mu = (I - P_m) P \mu = (I - P_m) G \beta^T$$

$$= (I - P_m) \sum_{i=1}^{p_m} \beta_i^T \bar{g}_i = \sum_{i \notin m} \beta_i^T (I - P_m) \bar{g}_i .$$

Since the vectors $\bar{g}_i$ are assumed to be linearly independent and any $P_m \bar{g}_i$ is an element of the column space of $G_m$, this shows that (A1) is equivalent to $\beta_i^T = 0$ for all $i \notin m$. The statements of the lemma are then obvious. \hfill \Box
Proof of Proposition 2.1. (i) Let \( \hat{m} \) be an arbitrary \( M_n^\ast \)-consistent procedure.

Then, given any \( \eta > 0 \), we obtain

\[
P \left\{ \left| \frac{D_n(\hat{m})}{d_n} - 1 \right| > \eta \right\} = P \left\{ \left| \frac{D_n(\hat{m})}{d_n} - 1 \right| > \eta, \hat{m} \in M_n^\ast \right\} + P \left\{ \left| \frac{D_n(\hat{m})}{d_n} - 1 \right| > \eta, \hat{m} \notin M_n^\ast \right\} \leq P(\hat{m} \notin M_n^\ast),
\]

which converges to zero because of the \( M_n^\ast \)-consistency of \( \hat{m} \).

(ii) Assume now that \( \hat{m} \) is asymptotically optimal and \( P(\hat{m} \in M_n^0) \to 1 \) as \( n \to \infty \). Thus it remains to show that \( P(\hat{m} \in M_n^0 \setminus M_n^\ast) \to 0 \).

Condition (2.2) provides firstly that for any \( \epsilon > 0 \) there exist \( n_0 \) and \( M_\epsilon \) such that \( P(d_n/d_n^\ast > M_\epsilon) < \epsilon/2 \) for all \( n \geq n_0 \).

Given any \( \epsilon > 0 \), let \( \eta_\epsilon = 1/M_\epsilon \). Then the asymptotic optimality of \( \hat{m} \) ensures that there is some \( n^\ast \geq n_0 \) such that \( P(|D_n(\hat{m})/d_n - 1| > \eta_\epsilon) < \epsilon/2 \) for all \( n \geq n^\ast \).

Consequently, we obtain for all \( n \geq n^\ast \)

\[
P(\hat{m} \in M_n^0 \setminus M_n^\ast) = P(\hat{m} \in M_n^0 \setminus M_n^\ast, |D_n(\hat{m})/d_n - 1| > \eta_\epsilon)
\]

\[
+ P(\hat{m} \in M_n^0 \setminus M_n^\ast, |D_n(\hat{m})/d_n - 1| \leq \eta_\epsilon)
\]

\[
\leq P(|D_n(\hat{m})/d_n - 1| > \eta_\epsilon) + P(\inf_{m \in M_n^0 \setminus M_n^\ast} |D_n(m)/d_n - 1| \leq \eta_\epsilon)
\]

\[
< \epsilon/2 + P(d_n/d_n^\ast \geq M_\epsilon) < \epsilon,
\]

which finishes the proof. \( \square \)

Proof of Proposition 2.2. Because of \( P(\hat{m} \in M_n^0) \to 1 \) it remains to verify that \( P(\hat{m} \in M_n^0 \setminus M_n^\ast) \to 0 \) as \( n \to \infty \).

Now, on account of \( \rho_n(m, m^\ast) > 0 \) for all \( m \in M_n^0 \setminus M_n^\ast \),

\[
P(\hat{m} \in M_n^0 \setminus M_n^\ast) \leq P \left( \inf_{m \in M_n^0 \setminus M_n^\ast} C_n(m) - C_n(m^\ast) \leq 0 \right)
\]

\[
= P \left( \bigcup_{m \in M_n^0 \setminus M_n^\ast} \{ C_n(m) - C_n(m^\ast) - \rho_n(m, m^\ast) \leq -\rho_n(m, m^\ast) \} \right)
\]

\[
\leq P \left( \bigcup_{m \in M_n^0 \setminus M_n^\ast} \left\{ \left| \frac{C_n(m) - C_n(m^\ast) - \rho_n(m, m^\ast)}{\rho_n(m, m^\ast)} \right| \geq 1 \right\} \right)
\]

\[
= P \left( \sup_{m \in M_n^0 \setminus M_n^\ast} \left| \frac{C_n(m) - C_n(m^\ast)}{\rho_n(m, m^\ast)} - 1 \right| \geq 1 \right),
\]

which tends to zero as \( n \to \infty \) due to assumption (2.5). \( \square \)

Proof of Theorem 2.1. Recalling the definitions in Section 1, we observe first that

\[
RSS(m) = \frac{1}{n} \| \varepsilon \|^2 + L_n(m) + \frac{2}{n} \varepsilon^T (\mu - P_n y)
\]  
(A2)
and
\[
L_n(m) = \frac{1}{n} \|\mu - P_m y\|^2 = \Delta_n(m) + \frac{1}{n} \|P_m \varepsilon\|^2 .
\]  
(A3)

On account of (3.1) and \( y = \mu + \varepsilon \) we have thus, for any \( m, m' \in M_n \),
\[
C_n(m) - C_n(m') = \Delta_n(m) - \Delta_n(m') + \frac{1}{n} \varepsilon^T (P_{m'} - P_m) \varepsilon + \frac{2}{n} \varepsilon^T (P_{m'} - P_m) \mu
\]
\[+ h_n \frac{|m| - |m'|}{n} \sigma^2 . \]  
(A4)

As it is obvious from Example 2, the result may be shown by applying Proposition 2.2
with \( M_0^n = M_n, D_n(m) = \Delta_n(m), M^*_n = M_0, m^* = m_1 \) and \( \rho_n(m, m^*) = \Delta_n(m) - \delta_n \).

Then, using (A4) and the fact that \( P - P_m \) is nonnegative definite, we obtain
\[
\sup_{m \in M_n \setminus M_0} \left| \frac{C_n(m) - C_n(m_1)}{\Delta_n(m) - \delta_n} - 1 \right|
\]
\[= \sup_{m \in M_n \setminus M_0} \left| \frac{\varepsilon^T (P - P_m) \varepsilon + 2 \varepsilon^T (P - P_m) \mu + (|m| - p_m) h_n \sigma^2}{n(\Delta_n(m) - \delta_n)} \right|
\]
\[\leq [\varepsilon^T P \varepsilon + p_n h_n \sigma^2]/(n \delta_n^*) + 2 \sup_{m \in M_n \setminus M_0} \left| \frac{\varepsilon^T (P - P_m) \mu}{n(\Delta_n(m) - \delta_n)} \right| . \]  
(A5)

Because of assumption (2.7) we have \( p_n h_n = o(n \delta_n^*) \), and the Markov inequality gives, for any \( \eta > 0 \),
\[
P \left( \frac{\varepsilon^T P \varepsilon}{n \delta_n^*} > \eta \right) \leq \frac{E \varepsilon^T P \varepsilon}{\eta n \delta_n^*} = \frac{p_n \sigma^2}{\eta n \delta_n^*} ,
\]  
(A6)

which converges again by assumption to zero as \( n \to \infty \). Therefore, it remains to verify that the last term in (A5) converges in probability to zero. To accomplish this, we note first that
\[
\text{Var}[\varepsilon^T (P - P_m) \mu] = \| (P - P_m) \mu \|^2 \sigma^2 = n^2 (\Delta_n(m) - \delta_n) \sigma^2 .
\]

Consequently, given any \( \eta > 0 \), Chebychev’s inequality yields
\[
P \left( \sup_{m \in M_n \setminus M_0} \frac{\| \varepsilon^T (P - P_m) \mu \|}{n(\Delta_n(m) - \delta_n)} > \eta \right) \leq \sum_{m \in M_n \setminus M_0} P \left( \frac{\| \varepsilon^T (P - P_m) \mu \|}{n(\Delta_n(m) - \delta_n)} > \eta \right)
\]
\[\leq \sum_{m \in M_n \setminus M_0} \frac{n (\Delta_n(m) - \delta_n) \sigma^2}{\eta^2 n^2 (\Delta_n(m) - \delta_n)^2}
\]
\[= \frac{|M_n \setminus M_0| \sigma^2}{\eta^2 n \delta_n^*} , \]  
(A7)

which converges to zero again by assumption (2.7).  
\( \square \)

30
Proof of Corollary 2.1. We proceed as in the proof of Theorem 2.1, but with \( m^* = m_0 \). Because of \( m \subset m_0 \) for all \( m \in M_n^N \setminus M_0 \), we obtain then, instead of (A5), the following upper bound for the left side of (2.5):

\[
\frac{[\varepsilon^T P_{m_0} \varepsilon + |m_0| h_n \sigma^2]/(m \delta_n^*)}{n(M_n^N \setminus M_0)} = \frac{\varepsilon^T(P_{m_0} - P_m) \mu}{n(\Delta_n(m) - \delta_n)}.
\]

Observing \( |M_n^N \setminus M_0| \leq |m_0| \) and \( \text{Var}[\varepsilon^T(P_{m_0} - P_m) \mu] = n[\Delta_n(m) - \delta_n] \sigma^2 \), the \( M_0 \)-consistency of \( \hat{m} \) follows from (2.8) by the same arguments as in the previous proof. \( \square \)

Proof of Theorem 2.2. Because of assumption (2.7) and Theorem 2.1 it remains to show that

\[
P(\hat{m} \in M_0 \setminus \{m_0\}) \rightarrow 0 \text{ as } n \rightarrow \infty .
\] (A8)

This may be done by applying Proposition 2.2 with \( D_n(m) = R_n(m) \), \( M_0^0 = M_0 \) and hence \( M_n^* = \{m_0\} \) (see Example 1). Furthermore we use the specification \( \rho_n(m, m^*) = h_n(R_n(m) - R_n(m^*)) \) with \( m^* = m_0 \).

Consider now an arbitrary \( m \in M_0 \setminus \{m_0\} \). From Lemma 2.1 and its proof it is obvious that \( \Delta_n(m) = \Delta_n(m_0) = \delta_n \) and \( P_m \mu = P_{m_0} \mu \). On account of (A4) this leads to

\[
n[C_n(m) - C_n(m_0)] = \varepsilon^T(P_{m_0} - P_m) \varepsilon + h_n(|m| - |m_0|) \sigma^2
\]

and \( n \rho_n(m, m_0) = h_n(|m| - |m_0|) \sigma^2 \). Consequently, using \( |m| > |m_0| \) and \( |\varepsilon^T(P_{m_0} - P_m) \varepsilon| = \varepsilon^T(P_m - P_{m_0}) \varepsilon \), we obtain

\[
\sup_{m \in M_0 \setminus \{m_0\}} \left| \frac{C_n(m) - C_n(m_0)}{\rho_n(m, m_0)} - 1 \right| = \sup_{m \in M_0 \setminus \{m_0\}} \frac{\varepsilon^T(P_m - P_{m_0}) \varepsilon}{h_n(|m| - |m_0|) \sigma^2} .
\] (A9)

First we observe that (A9) is bounded from above by \( \varepsilon^T(P - P_{m_0}) \varepsilon / h_n \) since \( \varepsilon^T(P_m - P_{m_0}) \varepsilon \leq \varepsilon^T(P - P_{m_0}) \varepsilon \) and \( |m| - |m_0| \geq 1 \) for all \( m \in M_0 \setminus \{m_0\} \). Thus, the Markov inequality provides, for any \( \eta > 0 \),

\[
P\left( \sup_{m \in M_0 \setminus \{m_0\}} \frac{\varepsilon^T(P_m - P_{m_0}) \varepsilon}{h_n(|m| - |m_0|) \sigma^2} > \eta \right) \leq \frac{(p_n - |m_0|)}{\eta h_n \sigma^2} .
\] (A10)

On the other hand, (A9) may also be used together with the Markov inequality to establish (2.5) as follows:

\[
P\left( \sup_{m \in M_0 \setminus \{m_0\}} \frac{\varepsilon^T(P_m - P_{m_0}) \varepsilon}{h_n(|m| - |m_0|) \sigma^2} > \eta \right) \leq \sum_{m \in M_0 \setminus \{m_0\}} P\left( \frac{\varepsilon^T(P_m - P_{m_0}) \varepsilon}{h_n(|m| - |m_0|) \sigma^2} > \eta \right)
\]

\[
\leq \sum_{m \in M_0 \setminus \{m_0\}} \frac{(|m| - |m_0|) \sigma^2}{\eta h_n(|m| - |m_0|) \sigma^2} \leq \frac{|M_0|}{\eta h_n} .
\] (A11)
Finally, in view of Proposition 2.2 the result is a consequence of (A10), (A11) and assumption (2.11).

**Proof of Corollary 2.2.** In the proof of Theorem 2.2, an upper bound for (A9) is obviously given by

\[
\sup_{m \in M_0 \setminus \{m_0\}} \frac{\varepsilon^T (P_m - P_{m_0}) \varepsilon - (|m| - |m_0|) \sigma^2}{h_n(|m| - |m_0|) \sigma^2} + \frac{1}{h_n}.
\]

It is therefore enough to show that the first term converges in probability to zero under \( h_n \to \infty \) as \( n \to \infty \).

Given any \( \eta > 0 \), we obtain for some constant \( C > 0 \)

\[
P \left( \sup_{m \in M_0 \setminus \{m_0\}} \frac{\varepsilon^T (P_m - P_{m_0}) \varepsilon - (|m| - |m_0|) \sigma^2}{h_n(|m| - |m_0|) \sigma^2} > \eta \right) \leq \sum_{m \in M_0 \setminus \{m_0\}} \frac{E(\varepsilon^T (P_m - P_{m_0}) \varepsilon - (|m| - |m_0|) \sigma^2)^4}{\eta^4 h_n^4(|m| - |m_0|)^4 \sigma^8} \leq C \sum_{m \in M_0 \setminus \{m_0\}} \frac{\text{tr}(P_m - P_{m_0})^2)^2}{h_n^4(|m| - |m_0|)^4},
\]

(A12)

where the Markov inequality and Theorem 2 of Whittle (1960), respectively, have been used. For \( m \in M_0 \setminus \{m_0\} \) it holds \( m_0 \subseteq m \) and thus \( \text{tr}(P_m - P_{m_0})^2 = \text{tr}(P_m - P_{m_0}) = |m| - |m_0| \). In the case of nested model candidates, (A12) can therefore be rewritten as

\[
C h_n^{-4} \sum_{m \in M_0 \setminus \{m_0\}} (|m| - |m_0|)^{-2} = C h_n^{-4} \sum_{i=1}^{p_n - |m_0|} i^{-2},
\]

which converges to zero as \( n \to \infty \), since \( \sum_{i=1}^{\infty} i^{-2} \) is convergent and \( h_n \to \infty \) was assumed. \( \square \)

**Proof of Theorem 2.3.** Let \( \hat{m} \) be an arbitrary \( m_0 \)-consistent model selection procedure.

For any \( m \in M_0 \) we have \( \Delta_n(m) = \delta_n \) and hence, on account of (A3),

\[
L_n(m) = \delta_n + n^{-1} \|P_m \varepsilon\|^2 = \delta_n + n^{-1} \|P_m - P_{m_0}\| \varepsilon\|^2 + n^{-1} \|P_{m_0} \varepsilon\|^2 \geq L_n(m_0), \quad (A13)
\]

which gives \( m_0 \in M_n^*(L_n, M_0) \) and thus the \( M_n^*(L_n, M_0) \)-consistency of \( \hat{m} \).

Decomposing the model bias provides

\[
nR_n(m) = n\delta_n + \|P - P_m\| \mu\|^2 + |m| \sigma^2 = nR_n(m_0) + \|P - P_m\| \mu\|^2 + (|m| - |m_0|) \sigma^2.
\]

32
Recalling \( \| (P - P_n)\mu \|^2 \geq n\delta_n^* \) for all \( m \notin M_0 \), assumption (2.10) establishes then that \( m_0 \) is the unique minimizer of the risk \( R_n(m) \) for sufficiently large \( n \). That is, we obtain \( M^*_n = \{ m_0 \} \) for sufficiently large \( n \), which leads to the \( M^*_n \)-consistency of \( \hat{m} \).

Finally, the \( M^*_n \)-consistency of \( \hat{m} \) will follow by showing that \( P(m_0 \in M^*_n) \to 1 \) as \( n \to \infty \). Clearly, \( L_n(m) \geq \delta_n + \delta_n^* \) for all \( m \notin M_0 \). Invoking the Markov inequality yields therefore, on account of (A13),

\[
P(m_0 \notin M^*_n) = P(L_n(m_0) > \min_{m \in M_n \setminus M_0} L_n(m)) \\
\leq P(L_n(m_0) > \delta_n + \delta_n^*) \\
= P(\|P_0 \varepsilon\|^2 > n\delta_n^*) \leq |m_0|\sigma^2/(n\delta_n^*) ,
\]

which converges to 0 by (2.10) and entails the result. \( \square \)

**Proof of Proposition 2.3.** (i) Given any \( \eta > 0 \), set \( \gamma = \eta/(2+\eta) \). Then, recalling the definition of \( \hat{m} \), we obtain

\[
P \left\{ \left| \frac{D_n(\hat{m})}{d_n} - 1 \right| > \eta \right\} = P \left\{ \frac{D_n(\hat{m})}{d_n} > \frac{1+\gamma}{1-\gamma} \right\} \\
= P\{ (1 - \gamma)D_n(\hat{m}) - (1 + \gamma)d_n > 0 \} \\
\leq P\{ (1 - \gamma)D_n(\hat{m}) - (1 + \gamma)d_n > s_n[C_n(\hat{m}) - C_n(m^*)] \} \\
\leq P\{ [D_n(\hat{m}) - d_n - s_n[C_n(\hat{m}) - C_n(m^*)]] > \gamma D_n(\hat{m}) + d_n \} \\
\leq P \left\{ \sup_{m \in M_n} \frac{s_n[C_n(m) - C_n(m^*)] - [D_n(\hat{m}) - d_n]}{D_n(m) + d_n} > \gamma \right\},
\]

which converges to zero due to assumption (2.12).

(ii) If \( C_n(m) \) is defined by (2.6), then \( C_n(m) - C_n(m^*) = Z_n(m) - Z_n(m^*) \), compare (A2) and (2.14). Hence, for \( s_n = 1 \), the left side of (2.12) is bounded from above by twice the left side of (2.13), completing the proof. \( \square \)

**Proof of Theorem 2.4.** (i) **Asymptotic \( r_n \)-optimality of \( \hat{m} \).** In order to apply part (ii) of Proposition 2.3 with \( D_n(m) = R_n(m) \), we observe first that, on account of (1.4), (A3) and (2.14),

\[
n[Z_n(m) - R_n(m)] = 2\varepsilon^T (I - P_m)\mu - \varepsilon^T P_m \varepsilon + (h_n - 1)|m|\sigma^2 .
\]

Consequently, in view of (2.13) and (2.17), it suffices to show that, as \( n \to \infty \),

\[
\sup_{m \in M_n} \left| \frac{\varepsilon^T P_m \varepsilon - |m|\sigma^2}{nR_n(m)} \right| \xrightarrow{P} 0 \quad \text{and} \quad (A14)
\]

\[
\sup_{m \in M_n} \left| \frac{\varepsilon^T (I - P_m)\mu}{nR_n(m)} \right| \xrightarrow{P} 0 . \quad (A15)
\]
Both (A14) and (A15) can now be established using the Markov inequality and Theorem 2 of Whittle (1960).

Given any $\eta > 0$, we have for some positive constants $C$ and $\tilde{C}$

$$P\left(\sup_{m \in M_n} \left| \frac{e^T P_m \epsilon - |m| \sigma^2}{n R_n(m)} \right| > \eta \right) \leq \sum_{m \in M_n} P\left(\left| \frac{e^T P_m \epsilon - |m| \sigma^2}{n R_n(m)} \right| > \eta \right)$$

$$\leq \sum_{m \in M_n} \frac{E|e^T P_m \epsilon - |m| \sigma^2|^{2q}}{\eta^{2q} n^{2q} R_n(m)^{2q}}$$

$$\leq C \sum_{m \in M_n} \frac{|m|^q}{n^{2q} R_n(m)^{2q}} \leq \tilde{C} \sum_{m \in M_n} [n R_n(m)]^{-q},$$

which tends to zero as $n \to \infty$ due to assumption (2.16). Notice that the last inequality is a consequence of $|m| \sigma^2 \leq n R_n(m)$. The convergence (A15) follows by the same arguments, noting that, for some positive constant $C'$,

$$E|e^T (I - P_m) \mu|^{2q} \leq C'[\mu^T (I - P_m) \mu]^{q} = C'n \Delta_n(m)^{q},$$

and $\Delta_n(m) \leq R_n(m)$ for all $m \in M_n$.

(ii) Asymptotic $l_n$-optimality of $\hat{m}$. Because of $n[L_n(m) - R_n(m)] = e^T P_m \epsilon - |m| \sigma^2$, (A14) provides

$$\sup_{m \in M_n} \left| \frac{L_n(m)}{R_n(m)} - 1 \right| \xrightarrow{P} 0 \text{ as } n \to \infty. \quad (A16)$$

Consequently, part (i) of the proof leads to

$$\sup_{m \in M_n} \left| \frac{Z_n(m) - L_n(m)}{R_n(m)} \right| \xrightarrow{P} 0 \text{ as } n \to \infty. \quad (A17)$$

Now, given any $\eta > 0$, we obtain

$$P\left(\sup_{m \in M_n} \left| \frac{Z_n(m)}{L_n(m)} - 1 \right| > \eta \right)$$

$$\leq P\left(\sup_{m \in M_n} \left| \frac{Z_n(m)}{L_n(m)} - 1 \right| > \eta, \sup_{m \in M_n} \left| \frac{L_n(m)}{R_n(m)} - 1 \right| \leq \frac{1}{2} \right) + P\left(\sup_{m \in M_n} \left| \frac{L_n(m)}{R_n(m)} - 1 \right| > \frac{1}{2} \right)$$

$$\leq P\left(\sup_{m \in M_n} \left| \frac{Z_n(m) - L_n(m)}{R_n(m)} \right| > \frac{\eta}{2} \right) + P\left(\sup_{m \in M_n} \left| \frac{L_n(m)}{R_n(m)} - 1 \right| > \frac{1}{2} \right),$$

which tends to zero because of (A14) and (A15). Finally, the asymptotic $l_n$-optimality of $\hat{m}$ follows by part (ii) of Proposition 2.3.

\[\square\]

**Proof of Corollary 2.4.** In view of Theorem 2.4 it remains to verify that (2.16) holds with $q = 2$.
The assumption $nr_n \to \infty$ ensures first the existence of some slowly increasing sequence \{\(k_n\)\} of natural numbers such that $k_n \to \infty$ and $k_n = o([nr_n]^2)$ as $n \to \infty$. Then, on account of $M_n = M_n^n$, $R_n(m) \geq r_n$ and $nR_n(m) \geq |m|\sigma^2$, we obtain

$$\sum_{m \in M_n} [nR_n(m)]^{-2} \leq k_n(nr_n)^{-2} + \sigma^{-4} \sum_{|m|=k_n} |m|^{-2},$$

which converges to zero by the above choice of $k_n$ and the fact that $\sum_{i=1}^{\infty} i^{-2}$ is convergent.

**Proof of results in Subsection 2.4.** Under the orthonormality assumption (2.20), we obtain

$$R_n(m) = \delta_n + \frac{\sigma^2}{n} \sum_{j=1}^{p_n} \tau_j^2 + \frac{\sigma^2}{n} \sum_{j \in m} (1 - \tau_j^2),$$

leading to

$$m_r^* := \arg \min_{m \subseteq m_1} R_n(m) = \{j \in m_1 \mid \tau_j^2 > 1\},$$

$$r_n = R_n(m_r^*) = \delta_n + \frac{\sigma^2}{n} \sum_{j=1}^{p_n} \{\tau_j^2 I(\tau_j^2 \leq 1) + I(\tau_j^2 > 1)\}.$$

In the adequate case we have $\delta_n = 0$ and, for sufficiently large $n$, $r_n = \sigma^2 k_0/n$, since then $j \in \{1, \ldots, k_0\}$ if and only if $\tau_j^2 > 1$.

Similarly, we get

$$L_n(\hat{m}) = \delta_n + \frac{\sigma^2}{n} \sum_{j=1}^{p_n} \{\tau_j^2 I(\hat{\tau}_j^2 \leq h_n) + (\hat{\tau}_j - \tau_j)^2 I(\hat{\tau}_j^2 > h_n)\},$$

providing the “overall risk”

$$EL_n(\hat{m}) = \delta_n + \frac{\sigma^2}{n} \sum_{j=1}^{p_n} B(h_n, |\tau_j|),$$

where, as in Droge (1993),

$$B(a^2, \tau) = \tau^2 \{\Phi(a - \tau) + \Phi(a + \tau) - 1\} + (a + \tau)\varphi(a + \tau) + (a - \tau)\varphi(a - \tau) - \Phi(a + \tau) - \Phi(a - \tau) + 2 = 1 - F_3(a^2; \tau^2) - \tau^2 F_5(a^2; \tau^2) + 2\tau^2 F_3(a^2; \tau^2).$$

Consequently, with $\text{reg}(h_n, \tau) = B(h_n, \tau) - \min(\tau^2, 1)$, we see that

$$\text{Reg}(\hat{f}, \hat{m}) := EL_n(\hat{m}) - r_n = \frac{\sigma^2}{n} \sum_{j=1}^{p_n} \text{reg}(h_n, |\tau_j|)$$

35
is just the regret risk of \( \hat{m} \), cf. Droge (1993), which has to vanish compared with the minimal risk \( \tau_n \) as \( n \to \infty \) for asymptotically mean efficient procedures. In the adequate case know that \( \tau_j = 0 \) unless \( j \in \{i_1, \ldots, i_{k_0}\} \), so that for sufficiently large \( n \)

\[
\frac{n}{\sigma^2} \text{Reg}(f, \hat{m}) = (p_n - k_0)B(h_n, 0) + \sum_{j=1}^{k_0} B(h_n, |\tau_{ij}|) - 1
\]

Observing

\[
B(h_n, 0) = 2\left[\Phi(-h_n^{1/2}) + h_n^{1/2}\varphi(h_n^{1/2})\right] \leq 2\varphi(h_n^{1/2}) (h_n^{-1/2} + h_n^{1/2})
\]

and, on account of (2.25) for \( k = 3 \),

\[
0 \leq \sum_{j=1}^{k_0} (B(h_n, |\tau_{ij}|) - 1) \leq 2 \sum_{j=1}^{k_0} \tau_{ij}^2 F_3(h_n; \tau_{ij}^2) \leq 3^{-3/2} 2 \sum_{j=1}^{k_0} \tau_{ij}^2 \exp(h_n - \tau_{ij}^2 / 3),
\]

we obtain, for some constant \( C \),

\[
\frac{n}{\sigma^2} \text{Reg}(f, \hat{m}) \leq C \exp\{\ln(p_n) + \ln(1 + h_n) - \ln(h_n)/2 + h_n/2 + \ln(n\delta_n^*) - (n\delta_n^*)/3\},
\]

which converges to zero if \( h_n \to \infty \), \( h_n/n \to 0 \) and \( \ln(p_n)/h_n \to 0 \) (recall that \( \liminf_{n \to \infty} \delta_n^* > 0 \)). Under these conditions the procedure \( \hat{m} \) is therefore asymptotically mean efficient, since \( n\tau_n = k_0\sigma^2 \) for sufficiently large \( n \).

In case of a bounded penalty sequence \( h_n \), we know from Droge (1993) that the “individual regret function” \( \text{reg}(h_n, \tau) \) is bounded. Therefore, a simple sufficient condition for achieving asymptotic mean efficiency is \( p_n/(nr_n) \to 0 \), which requires, of course, that all model candidates are inadequate. Roughly speaking, the decision whether procedures with small (bounded) or large (“tending to \( \infty \”) penalttaries are preferable depends on the portion of standardized coefficients (\( \tau_i \)) tending to \( \infty \) and zero, respectively, (and their rates). This is due to the fact that, for bounded \( h_n \), \( \lim_{|\tau| \to \infty} \text{reg}(h_n, \tau) = 0 \), whereas \( \text{reg}(h_n, 0) \to 0 \) is only possible if \( h_n \to \infty \), which in turn implies \( \sup_{\tau} \text{reg}(h_n, \tau) \to \infty \), see Droge (1993).

Proof of Lemma 3.1. Observing

\[
\hat{\sigma}_n^2(m) = t_m^{-1} \|\mu\|_{T_m}^2 + t_m^{-1} \|\varepsilon\|_{T_m}^2 + 2t_m^{-1} \mu^T T_m \varepsilon, \quad (A18)
\]

we obtain statement (i) because of \( E\varepsilon = 0 \) and \( E\|\varepsilon\|_{T_m}^2 = t_m \sigma^2 \).

Finally, statement (ii) follows by using (A18), statement (i), and Theorem 2 of Whittle (1960); that is we have for some positive constants \( C_1 \) and \( C_2 \),

\[
E[\hat{\sigma}_n^2(m) - \varepsilon_n(m)]^{2q} = E[t_m^{-1} \|\varepsilon\|_{T_m}^2 - \sigma^2 + 2t_m^{-1} \mu^T T_m \varepsilon]^{2q}
\]

36
\[
\begin{align*}
&\leq 2^{2t-1}\left\{ E\|t_m^{-1}\|\varepsilon\|T_m - \sigma^2\|^2 + E\|2t_m^{-1}\mu^T T_m \varepsilon\|^2 \right\} \\
&\leq C_1 t_m^{-2q}[\text{tr}(T_m^2)]^q + C_2 t_m^{-2q}[\mu^T T_m^2 \mu]^q \\
&\leq C_1 t_m^{-2q}(\omega(m))^q + C_2 t_m^{-2q}[\omega(m)\|\mu\|^2]^{2q} = O\left([t_m^{-1}\omega(m)\epsilon_n(m)]^q\right).
\end{align*}
\]

For the last inequality we have used the fact that \( \omega(m)T_m - T_m^2 \) is nonnegative definite and thus \( \text{tr}(T_m^2) \leq \omega(m)t_m \).

**Proof of Theorem 3.1.** (i) In view of Proposition 2.2 and the proof of Theorem 2.1, the \( M_0 \)-consistency of a procedure minimizing (3.1) over \( m \in M_n \) (general case (a)) follows by showing that

\[
\sup_{m \in M_n \setminus M_0} \frac{\|h_n(m)|m| - \hat{h}_n(m_1)p_n\|\sigma^2}{n(\Delta_n(m) - \delta_n)} = O_P(1) \quad \text{as} \quad n \to \infty.
\]  

(A19)

Therefore it suffices to establish that, as \( n \to \infty \),

\[
A_n := \sup_{m \in M_n \setminus M_0} \frac{\hat{h}_n(m)|m|\sigma^2}{n(\Delta_n(m) - \delta_n)} = o_P(1) \quad \text{and} \quad (A20)
\]

\[
B_n := \frac{\hat{h}_n(m_1)p_n\sigma^2}{n\delta_n^*} = o_P(1) .
\]  

(A21)

In both cases (a) and (b), we have \( \hat{h}_n(m_1) = h_n(m_1)\hat{\sigma}^2(m_1)/\sigma^2 \). On account of (2.7) and the definition of \( h_n \), we obtain therefore

\[
EB_n \leq h_n p_n/(n\delta_n^*) \to 0 \quad \text{as} \quad n \to \infty ,
\]

so that the Markov inequality provides (A21).

Under (a) it holds, for any \( m \in M_n \), \( \hat{h}_n(m) \sigma^2 = h_n(m)\hat{\sigma}^2(m_1) \leq h_n\hat{\sigma}^2(m_1) \), so that (A20) is obtained by the same arguments as used for (A21), recall \( \Delta_n(m) - \delta_n \geq \delta_n^* \).

Consider now case (b), where \( \hat{h}_n(m_1) \sigma^2 = h_n(m_1)\hat{\sigma}^2(m) = h_n(m)n(n - |m|)^{-1}RSS(m) \).

Using (A2) and (A3), we obtain

\[
RSS(m) - RSS(m_1) = \Delta_n(m) - \delta_n + \frac{1}{n}\varepsilon^T(P - P_m)\varepsilon + \frac{2}{n}\varepsilon^T(P - P_m)\mu ,
\]

and, consequently, \( \hat{h}_n(m)|m|\sigma^2 \leq \hat{h}_n p_n(n - p_n^{-1} \{ RSS(m) - RSS(m_1) \} + RSS(m_1)) \).

Because of the nonnegative definiteness of \( P - P_m \), this gives

\[
A_n \leq \frac{\hat{h}_n p_n}{n - p_n} \left( 1 + \frac{\varepsilon^T P \varepsilon}{n\delta_n^*} + 2 \sup_{m \in M_n \setminus M_0} \frac{\varepsilon^T(P - P_m)\mu^2}{n(\Delta_n(m) - \delta_n)} \right) + \frac{\hat{h}_n p_n \hat{\sigma}^2(m_1)}{n\delta_n^*} ,
\]  

(A22)

where the last term converges in probability to zero as proved above. Finally, the first summand on the right hand side of (A22) converges also in probability to zero due to (A6), (A7) and the assumption made under (b).
(ii) In the special case of nested model candidates, the proof follows the lines of the proof of Corollary 2.1. That is, we take \( m^* = m_0 \) instead of \( m^* = m_1 \), and use the fact that \( m \subset m_0 \) for all \( m \in M_n^N \setminus M_0 \). Consequently, it remains to verify (A19), or (A20) and (A21), where \( M_n, m_1 \) and \( p_n \) have to be replaced by \( M_n^N, m_0 \) and \( \|m_0]\), respectively.

Under (a) we observe \( \hat{\sigma}_n^2(m_0) = \hat{\sigma}^2(m_1) \) and thus \( \hat{h}_n(m_0) \leq \hat{h}_n^0(\hat{\sigma}^2(m_1)) \), so that both \( E A_n \) and \( E B_n \) are bounded from above by \( \hat{h}_n^0(\|m_0|E \hat{\sigma}^2(m_1) / \|m_0|)/(n \delta_n) = h_n(m_0)/(n \delta_n) \). This bound tends to zero due to (2.8), leading to the result by the Markov inequality.

Under (b) it holds \( \hat{h}_n(m_0) \leq \hat{h}_n^0(\hat{\sigma}^2(m_0)) \). Then we proceed as in the general case (i), but with the already mentioned replacements and with \( P_{m_0} \) instead of \( P \) (e.g. in (A22)), and obtain the desired property on account of the modified assumptions. \( \square \)

**Proof of Corollary 3.1.** For the generalized \( C_p \)-criterion we have obviously \( \tilde{h}_n = \kappa_n \). Table 3.1 shows furthermore that \( \tilde{h}_n \leq 2 \) holds for each of the criteria \( C_p, FPE, SH, GCV \) and GFCV. Consequently, the assumptions of the corollary imply that the assumptions of Theorem 3.1 are satisfied, so that the \( M_0 \)-consistency of the corresponding procedures follows.

To obtain the same for GIC (and AIC as special case), it is enough to verify that \( \tilde{h}_n/a_n \to 1 \) as \( n \to \infty \), where \( \tilde{h}_n := \sup_{m \in M_n \setminus M_0} h_n(m) \), see the proof of Theorem 3.1. We note first that \( q_n^*(m) \leq q_n(m_1) =: q_n \) for all \( m \in M_n \) (or, in case (iii), \( q_n(m) \leq q_n(m_0) =: q_n \) for all \( m \in M_n^N \setminus M_0 \). The assumption (3.7) (or, in case (iii), \( a_n|m_0|/n \to 0 \)) ensures now that always \( q_n \to 1 \) as \( n \to \infty \), and hence \( \lim_{n \to \infty} \tilde{h}_n/a_n = \lim_{n \to \infty} a_n q_n/a_n = 1 \). \( \square \)

**Proof of Theorem 3.2.** As in the proof of Theorem 3.1, it suffices again to establish (A20) and (A21). We use the notation of Table 3.1 and Lemma 3.1.

Let \( \omega_n := \sup_{m \in M_n} \omega_n(m) \). Then we obtain, for any \( m \in M_n \),

\[
\hat{h}_n(m)|m|\sigma^2 = \|I - P_m\|\sigma^2 \leq \omega_n RSS(m),
\]

(A23)\)

where \( \Omega_m = \Gamma(m), \omega_n = c_n(2 - c_n)(1 - c_n)^{-2} \) (for CV) or \( \Omega_m = \Lambda(m), \omega_n = c_n(2 + c_n) \) (for FCV). Because of (3.10) we have, for both criteria, \( \omega_n = 2c_n(1 + o(1)) \) as \( n \to \infty \).

Using additionally the obvious relation \( c_n \geq p_n/n \), we get (3.6). As before, (A21) is therefore a consequence of the Markov inequality, since

\[
EB_n \leq \frac{\omega_n(n - p_n)E\hat{\sigma}^2(m_1)}{n\delta_n} = \frac{2c_n(1 + o(1))(n - p_n)(\sigma^2 + n(n - p_n)^{-1}\delta_n)}{n\delta_n},
\]

which tends to zero by (3.10), (3.11) and (3.6).
To prove (A20), we apply first inequality (A23) and proceed then as in the proof of Theorem 3.1. This leads to (A22), but with \( \omega_n \) instead of \( \bar{h}_n p_n/(n - p_n) \), so that the result follows by \( \omega_n \to 0 \) as \( n \to \infty \).

Finally, the required modifications in the special case \( M_n = M_n^N \) are obvious from the proof of Theorem 3.1: the pseudo-true model \( m_0 \) plays now the role of the largest model \( m_1 \), and any pseudo-inadequate model is necessarily a submodel of \( m_0 \). \( \square \)

**Proof of Theorem 3.3.** In view of Corollary 3.1 (ii), the \( M_0 \)-consistency of the procedure follows if we can establish condition (3.4), where \( \bar{h}_n \) is replaced by \( a_n \). Because of (2.11) we have \( a_n \to \infty \) as \( n \to \infty \), so that (3.7) leads to (3.6). Hence, on account of (3.14), condition (2.7) provides (3.4).

Note that the assumptions ensure the (weak) asymptotic equivalence of \( a_n, h_n \) and \( \bar{h}_n \), where the last two quantities are defined by (3.13) and (3.3), respectively. This is easily seen by the above considerations and the observation that \( q_n = q_n(m_1) \to 1 \) as \( n \to \infty \) due to condition (3.7). Consequently, conditions (2.7), (2.11), (3.7) and (3.4) are satisfied for each of the three quantities \( a_n, h_n \) and \( \bar{h}_n \).

To establish the \( m_0 \)-consistency, we proceed as in Theorem 2.2, but now with

\[
\rho_n(m, m_0) = n^{-1} a_n q_n(m) |m| - q_n(m_0) |m_0| \sigma^2 + \delta_n \, .
\]

This choice fulfills the requirement of Proposition 2.2 since, for any \( m \in M_0 \setminus \{m_0\} \), we get \( |m| > |m_0|, 1 \leq q_n(m_0) \leq q_n(m) \leq q_n \) and therefore

\[
n p_n(m, m_0) \geq a_n q_n(|m| - |m_0|) \sigma^2 + \delta_n \geq h_n(|m| - |m_0|) \sigma^2 > 0 \, ,
\]

see (3.13) for the definition of \( h_n \). According to the proof of Theorem 2.2, the following convergences have to be verified, as \( n \to \infty \):

\[
\sup_{m \in M_0 \setminus \{m_0\}} \frac{\varepsilon^T (P_m - P_{m_0}) \varepsilon}{n \rho_n(m, m_0)} = o_P(1) \quad \text{and} \quad (A25)
\]

\[
\sup_{m \in M_0 \setminus \{m_0\}} \left| \frac{(\hat{h}_n(m)|m| - \hat{h}_n(m_0)|m_0|) \sigma^2}{n \rho_n(m, m_0)} \right| - 1 = o_P(1) \, .
\]

Because of (A24), the left hand side of (A25) can be estimated as in (A10) and (A11), so that the first result follows by assumption (2.11). Concerning (A26) we note that

\[
R S S(m) - R S S(m_0) = -n^{-1} \varepsilon^T (P_m - P_{m_0}) \varepsilon \quad \text{for all} \quad m \in M_0 \setminus \{m_0\},
\]

which leads to

\[
(\hat{h}_n(m)|m| - \hat{h}_n(m_0)|m_0|) \sigma^2 = -n^{-1} a_n q_n(m)|m| \varepsilon^T (P_m - P_{m_0}) \varepsilon + a_n(q_n(m)|m| - q_n(m_0)|m_0|) R S S(m_0) \, .
\]

39
On account of $q_n(m)|m| \leq q_n p_n$ we obtain hence, for any $m \in M_0 \setminus \{m_0\}$,

$$\left| \frac{(\hat{h}_n(m)|m| - \hat{h}_n(m_0)|m_0|)\sigma^2}{np_n(m, m_0)} - 1 \right| \leq \frac{a_n q_n p_n}{n} \sup_{m \in M_0 \setminus \{m_0\}} \varepsilon^T(P_m - P_{m_0})\varepsilon + \frac{RSS(m_0)}{q_n(\sigma^2 + \delta_n)} - 1.$$

The first term on the right hand side converges in probability to zero due to (A25) and (3.7), so that it remains to show that the second term vanishes asymptotically, too. To accomplish this, we write, using (A2) and (A3),

$$RSS(m_0) - q_n(\sigma^2 + \delta_n) = (n^{-1}||\varepsilon||^2 - \sigma^2) - n^{-1}||P_m \varepsilon||^2 + 2n^{-1} \mu^T(I - P_m)\varepsilon - (q_n - 1)(\sigma^2 + \delta_n),$$

so that it suffices to prove that each of the four terms on the right hand side converges in probability to zero, recall that $q_n(\sigma^2 + \delta_n) \geq \sigma^2$. For the first term this holds due to the law of large numbers. Invoking Markov's and Chebychev's inequality, respectively, provides the result for the second and third term, since

$$En^{-1}||P_m \varepsilon||^2 = n^{-1}|m_0|\sigma^2 \leq n^{-1} p_n \sigma^2 \to 0 \quad \text{as} \quad n \to \infty,$$

compare (3.6), and

$$Var[n^{-1}\mu^T(I - P_m)\varepsilon] = E[n^{-1}\mu^T(I - P_m)\varepsilon]^2 = n^{-1} \sigma^2 \delta_n,$$

which tends to zero by (3.14). Finally, the last term tends to zero, since $q_n \to 1$ and $\delta_n = O(1)$ as $n \to \infty$. This completes the proof. $\Box$

**Proof of Corollary 3.2.** Conditions (3.15), (3.7) and $a_n \to \infty$ imply those versions of (2.8), (3.7) and (3.4), which are required in part (iii) of Corollary 3.1 to ensure the $M_0$-consistency of the procedure in case of nested model candidates.

To establish the $m_\sigma$-consistency of the procedure, we define $\rho_n(m, m_0)$ as in the proof of Theorem 3.3, and apply the idea of the proof of Corollary 2.2. For an arbitrary $m \in M_0 \setminus \{m_0\}$ we obtain then

$$\hat{C}_n(m) - \hat{C}_n(m_0) = \varepsilon^T(P_m - P_{m_0})\varepsilon + (\hat{h}_n(m)|m| - \hat{h}_n(m_0)|m_0|)\sigma^2
\neq (1 - n^{-1}a_n q_n(m)|m|)\varepsilon^T(P_m - P_{m_0})\varepsilon + a_n(q_n(m)|m| - q_n(m_0)|m_0|)RSS(m_0).$$

Using (A24), $q_n(m)|m| \leq q_n p_n$ and the triangular inequality, this leads to

$$\sup_{m \in M_0 \setminus \{m_0\}} \left| \frac{\hat{C}_n(m) - \hat{C}_n(m_0)}{\rho_n(m, m_0)} - 1 \right| \leq |\tau_n| S_1 + S_2 + S_3,$$

where

$$\tau_n = 1 + n^{-1}a_n q_n p_n, \quad S_1 = \sup_{m \in M_0 \setminus \{m_0\}} \frac{|\varepsilon^T(P_m - P_{m_0})\varepsilon - (|m| - |m_0|)\sigma^2|}{\hat{h}_n(|m| - |m_0|)\sigma^2},$$

$$S_2 = \frac{RSS(m_0)}{q_n(\sigma^2 + \delta_n)} - 1 \quad \text{and} \quad S_3 = \frac{\tau_n \sigma^2}{a_n q_n(\sigma^2 + \delta_n)},$$

40
Clearly, as \( n \to \infty \), we have \( \tau_n \to 1 \) by (3.7) and thus \( S_3 \to 0 \), since \( a_n \to \infty \). Moreover, from the proofs of Corollary 2.2 and Theorem 3.3, respectively, we know that \( S_1 \) and \( S_2 \) converge in probability to zero, completing the proof. \( \square \)

**Proof of Theorem 3.4.** (i) Clearly, condition (2.7) for \( h_n \) defined by (3.16) provides immediately conditions (2.7), with \( \kappa_n \) instead of \( h_n \), and (3.7) for \( h_n = \kappa_n \). This shows the \( M_0 \)-consistency of the procedure due to part (ii) of Corollary 3.1.

As in the proof of Theorem 3.3, to get the \( m_0 \)-consistency of the procedure it is enough to establish (A25) and (A26), where

\[
\rho_n(m, m_0) = n^{-1} h_n(\|m| - |m_0|)\sigma^2.
\]

The convergence (A25) follows again by (2.11). Recalling \( \hat{h}_n(m) = \kappa_n \hat{\sigma}^2(m_1)/\sigma^2 \), (A26) is equivalent to

\[
\left| \frac{\hat{\sigma}^2(m_1)}{E\hat{\sigma}^2(m_1)} - 1 \right| \to 0 \quad \text{as} \quad n \to \infty,
\]

which is in turn an easy consequence of Chebyshev’s inequality and Lemma 3.1, since \((n - p_n)\hat{\sigma}^2(m_1) \to \infty \) as \( n \to \infty \) by assumption (3.17).

(ii) Condition (2.8) for \( h_n \) defined by (3.16) leads now to (2.8), with \( \kappa_n \) instead of \( h_n \), \( \kappa_n m_0 \delta_n/[\|n - p_n = \delta^*_n\] \to 0 \) as \( n \to \infty \), and thus to the \( M_0 \)-consistency by Corollary 3.1 (iii).

To prove the \( m_0 \)-consistency, we proceed as in the proof of Corollary 2.2, again with \( h_n \) defined by (3.16), and find that besides the steps made there we need only to verify (A27). \( \square \)

**Proof of Proposition 3.1.** First we note that \( \hat{C}_n(m) - \hat{C}_n(m^*) = \hat{Z}_n(m) - \hat{Z}_n(m^*) \), where \( m^* \) is a minimizer of \( R_n(m) \) and \( \hat{Z}_n(m) \) is defined as \( Z_n(m) \) in (2.14) with \( \hat{h}_n(m) \) replacing \( h_n \); compare the proof of Proposition 2.3. In view of the proof of Theorem 2.4 it suffices therefore to establish (3.18), since (A14) and (A15) are implied by the assumptions (2.15) and (2.16). \( \square \)

**Proof of Theorem 3.5.** For the \( C_p \)-criterion we have \( \hat{h}_n(m) = \hat{h}_n = 2\hat{\sigma}^2(m_1)/\sigma^2 \), see (3.2) and Table 3.1. Let \( \xi_n = \sup_{m \in M_n} \frac{|m|}{nR_n(m)} \). In view of Proposition 3.1 it suffices then to show that, as \( n \to \infty \),

\[
|E\hat{\sigma}^2(m_1) - \sigma^2|\xi_n \to 0 \quad \text{and} \quad \left|E\hat{\sigma}^2(m_1) - E\hat{\sigma}^2(m_1)\right|^2\xi_n \to 0,
\]

from which (3.18) follows.
Lemma 3.1 yields now \(E \hat{\sigma}^2(m_1) - \sigma^2 \xi_n = \frac{n}{n-p_n} \delta_n \xi_n\) and, for some constant \(C > 0\),

\[
E|\hat{\sigma}^2(m_1) - E\hat{\sigma}^2(m_1)|^2 \xi_n^2 \leq C \left[ \frac{\sigma^2 \xi_n + \frac{n}{n-p_n} \delta_n \xi_n}{n-p_n} \right] \xi_n \leq C \left[ \frac{1 + \frac{n}{n-p_n} \delta_n \xi_n}{(n-p_n) \sigma^2} \right],
\]

since \(\xi_n \leq \sigma^{-2}\) due to \(n R_n(m) \geq |m| \sigma^2\). Noting that \(n - p_n \to \infty\) holds under each of the conditions (3.6), (3.19) and (3.20), it remains thus to verify that \(\frac{n}{n-p_n} \delta_n \xi_n \to 0\) as \(n \to \infty\). Under (3.20) this is obvious since \(\xi_n\) is bounded. On the other hand, using \(|m| \leq p_n\) and \(R_n(m) \geq r_n(\geq \delta_n)\), we obtain \(\frac{n}{n-p_n} \delta_n \xi_n \leq \frac{p_n}{(n-p_n)} r_n^{\delta_n}\), which tends to zero under both (3.6) and (3.19), completing the proof.

**Proof of Theorem 3.6.** Obviously, condition (3.18) and thus the result will follow by showing that, as \(n \to \infty\),

\[
b_n := \sup_{m \in M_n} \frac{|E \hat{h}_n(m) - 2|m| \sigma^2}{n R_n(m)} \to 0 \quad \text{and} \quad (A30)
\]

\[
v_n := \sup_{m \in M_n} \frac{||h_n(m) - E \hat{h}_n(m)|| \sigma^2}{n R_n(m)} \to 0. \quad (A31)
\]

Lemma 3.1 gives now \(E \hat{h}_n(m) = h_n(m)[\sigma^2 + \frac{n}{n-|m|} \Delta_n(m)]/\sigma^2\), leading to

\[
b_n = \sup_{m \in M_n} \frac{||h_n(m) - 2|m| \sigma^2 + \frac{n}{n-|m|} \Delta_n(m)||}{n R_n(m)}.
\]

Hence, (A30) is an immediate consequence of the triangular inequality, (3.22) and (3.23).

To establish (A31) we apply the Markov inequality and Lemma 3.1. Given any \(\eta > 0\), this yields for some constant \(C > 0\)

\[
P(v_n > \eta) \leq \sum_{m \in M_n} \frac{E|\hat{\sigma}^2(m) - E\hat{\sigma}^2(m)|^2 |m|^{2q} h_n(m) |m|^{2q}}{\eta^{2q} [n R_n(m)]^{2q}} \leq C \sum_{m \in M_n} \frac{|E\hat{\sigma}^2(m)|^2 |m|^{2q} h_n(m) |m|^{2q}}{[n R_n(m)]^{2q} (n-|m|)^q} \leq C \sum_{m \in M_n} \frac{\sigma^2 |m| + \frac{|m|}{n-|m|} |n \Delta_n(m)| |m|^{2q} h_n(m) |m|^{2q}}{[n R_n(m)]^{2q} (n-|m|)^q} \left( \frac{|m|}{n-|m|} \right)^q. \quad (A32)
\]

On account of (3.21) and (3.22), there exist now constants \(C_1 > 1\) and \(C_2 > 0\) such that for all \(n\)

\[
\frac{|m|}{n-|m|} \leq \frac{p_n}{n-p_n} \leq C_1 \quad \text{and} \quad \sup_{m \in M_n} h_n(m) \leq C_2.
\]

42
Hence, we have

$$\sigma^2 |m| + \frac{|m|}{n - |m|} n \Delta_n(m) \leq C_1 n R_n(m)$$

and (A32) does not exceed $C[C_1 C_2]^{2q} \sum_{m \in M_n} [n R_n(m)]^{-q}$, which in turn tends to zero by (2.16). This completes the proof. \hfill \Box

**Proof of Corollary 3.4.** According to the discussion before Corollary 3.4 it remains to verify that (3.24) holds for AIC, GCV, GFCV and SH. For the criteria SH, GCV and GFCV this is obvious, since then Table 3.1 yields that $\sup_{m \in M_n} |h_n(m) - 2|$ is given by $2p_n/n$, $p_n/(n - p_n)$ and $p_n/n + p_n^2/n^2$, respectively, leading to (3.24) by (3.6).

Consider now the AIC criterion, for which we have $h_n(m) = 2q_n(m)(1 - |m|/n)$ with $a_n = 2$, see Table 3.1. Straightforward calculations show that $h_n(m)$ is monotonically decreasing in $|m|$, with $h_n(m) \to 2$ if $|m|/n \to 0$. Consequently, we obtain for all $m \in M_n$

$$|h_n(m) - 2| = 2 - h_n(m) \leq 2 - h_n(m_1)$$

so that (3.24) follows again by (3.6). \hfill \Box

**Proof of Theorem 3.7.** Analogously to the proof of Theorem 3.6 it is enough to establish (A30) and (A31), but for the penalties $\hat{h}_n(m)$ associated with the criteria CV and FCV, which are given by (3.2) and Table 3.1.

Using the notations of Lemma 3.1, we get first

$$\hat{h}_n(m) |m| \sigma^2 = t_m \tilde{\sigma}_n^2(m), \quad E t_m \tilde{\sigma}_n^2(m) = t_m \sigma^2 + ||\mu||^2_{T_m}, \quad (A33)$$

and consequently

$$b_n \leq \sup_{m \in M_n} \frac{|t_m - 2|m|\sigma^2 + ||\mu||^2_{T_m}}{n R_n(m)}. \quad (A34)$$

Because of $\text{tr}(P_n) = |m|$ and the definition of $c_n$ we derive $|t_m - 2|m| \leq |m| \rho_n$, where $\rho_n$ is given for CV and FCV, respectively, by $c_n(1 - c_n)^{-1}$ and $c_n(1 + c_n)$. On account of $||\mu||^2_{T_m} \leq \omega(m)n \Delta_n(m)$, (A34) and (1.4), this leads to $b_n \leq \max\{\rho_n, \sup_{m \in M_n} \omega(m)\}$, which tends to zero for both CV and FCV due to assumption (3.10) and thus entails (A30).

To verify (A31) we apply (A33), the Markov inequality and Lemma 3.1. Given any $\eta > 0$, this yields for some constant $C > 0$

$$P(v_n > \eta) \leq \sum_{m \in M_n} \frac{E|\tilde{\sigma}_n^2(m) - E\tilde{\sigma}_n^2(m)|^2 q |t_m|^{2q}}{\eta^{2q}|n R_n(m)|^{2q}}$$

43
\[ \leq C \sum_{m \in M_n} \frac{|t_m E \delta_n^2(m)|^q \omega(m)^q}{[nR_n(m)]^{2q}} \]
\[ \leq CC_n \sum_{m \in M_n} [nR_n(m)]^{-q}, \] (A35)

where \( C_n = [\sup_{m \in M_n} \omega(m) \max\{2 + \rho_n, \omega(m)\}]^q \). The last inequality in (A35) results from
\[ En \delta_n^2(m) = t_m \sigma^2 + \|\mu\|^2 \leq (2 + \rho_n) m |\sigma^2 + \omega(m) n \Delta_n(m)| \leq C_n R_n(m), \]

compare the proof of (A30). Since \( C_n \) is bounded under condition (3.10) (it converges even to zero), (A35) tends to zero as \( n \to \infty \) by (2.16), completing the proof. \( \square \)

**Acknowledgement**

This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifizierung und Simulation ökonomischer Prozesse", Humboldt-Universität zu Berlin, Berlin, Germany.

**References**


