Model Selection

Hannes Leeb[1] and Benedikt M. Pötscher[2]

[1] Yale University hannes.leeb@yale.edu
[2] University of Vienna benedikt.poetscher@univie.ac.at

1 The Model Selection Problem

Model selection has become an ubiquitous statistical activity in the last decades, none the least due to the computational ease with which many statistical models can be fitted to data with the help of modern computing equipment. In this article we provide an introduction to the statistical aspects and implications of model selection and we review the relevant literature.

1.1 A General Formulation

When modeling data Y, a researcher often has available a menu of competing candidate models which could be used to describe the data. Let $\mathcal{M}$ denote the collection of these candidate models. Each model M, i.e., each element of $\mathcal{M}$, can – from a mathematical point of view – be viewed as a collection of probability distributions for Y implied by the model. That is, M is given by

$$M = \{\mathbb{P}_\eta : \eta \in H\},$$

where $\mathbb{P}_\eta$ denotes a probability distribution for Y and H represents the 'parameter' space (which can be different across different models M). The 'parameter' space H need not be finite-dimensional. Often, the 'parameter' η will be partitioned into (η_1, η_2) where η_1 is a finite-dimensional parameter whereas η_2 is infinite-dimensional. In case the parameterization is identified, i.e., the map $\eta \to \mathbb{P}_\eta$ is injective on H, we will often not distinguish between M and H and will use them synonymously.

The model selection problem is now to select – based on the data Y – a model $\hat{M} = \hat{M}(Y)$ in $\mathcal{M}$ such that $\hat{M}$ is a 'good' model for the data Y. Of course, the sense, in which the selected model should be a 'good' model, needs to be made precise and is a crucial point in the analysis. This is particularly important if – as is usually the case – selecting the model $\hat{M}$ is not the final

purpose of the analysis, but $\hat{M}$ is used as a basis for the construction of parameter estimators, predictors, or other inference procedures.

Typically, with each model M we will have associated an estimator $\hat{\eta}(M)$ such as the maximum likelihood estimator or the least squares estimator, etc. It is important to note that when model selection precedes parameter estimation, the estimator finally reported is $\tilde{\eta} = \hat{\eta}(\hat{M})$ (and *not* one of the estimators $\hat{\eta}(M)$). We call $\tilde{\eta}$ a post-model-selection estimator (PMSE). It is instructive to note that $\tilde{\eta}$ can be written as

$$\tilde{\eta} = \sum_{M \in \mathcal{M}} \hat{\eta}(M) \mathbf{1}(\hat{M} = M),$$

which clearly shows the compound nature of the PMSE. [Note that the above sum is well-defined even if the spaces H for different M bear no relationship to each other.]

We note that in the framework just described it may or may not be the case that one of the candidate models M in $\mathcal{M}$ is a correct model (in the sense that the actual distribution of the data coincides with a distribution $\mathbb{P}_\eta$ in M). A few examples illustrating the above notation are in order.

Example 1. (Selection of regressors) Suppose Y is an $n \times 1$ vector generated through

$$Y = X\theta + u \tag{1}$$

where X is an $n \times K$ matrix of non-stochastic regressors with full column-rank and u is a disturbance term whose distribution F does not depend on θ and varies in a set $\mathcal{F}$ of distributions (e.g., $\mathcal{F}$ could be the set of all $N(0, \sigma^2 I_n)$ distributions). Suppose the researcher suspects that some regressors (i.e. columns of X) are superfluous for explaining Y (in the sense that the true value of the coefficients of these regressors are zero), but does not know which of the regressors are superfluous. Then the appropriate candidate models are all submodels of (1) given by zero restrictions on the parameter vector θ. More formally, let $\mathfrak{r} \in \{0,1\}^K$, i.e., $\mathfrak{r}$ is a $K \times 1$ vector of zeros and ones. Then each $\mathfrak{r} \in \{0,1\}^K$ defines a submodel

$$M_{\mathfrak{r}} = \left\{ (\theta, F) \in \mathbb{R}^K \times \mathcal{F} : \theta_i = 0 \text{ if } \mathfrak{r}_i = 0 \right\},$$

the full model M_{full} corresponding to $\mathfrak{r} = (1, \ldots, 1)$. The set of all candidate models is given by

$$\mathcal{M}_{all} = \left\{ M_{\mathfrak{r}} : \mathfrak{r} \in \{0,1\}^K \right\}.$$

The set-up just described could be termed 'all-subset selection'. If – on a priori grounds – one wants to protect some of the variables, say the first k ones, from being eliminated by the model selection procedure, one would then of course consider as candidate models only those in the set

$$\mathcal{M}_{protected} = \left\{ M_{\mathfrak{r}} : \mathfrak{r} \in \{0,1\}^K,\ \mathfrak{r}_i = 1 \text{ for } i = 1, \ldots k \right\}.$$

Another case arises if there is an a priori given ordering of the regressors reflecting their perceived 'importance' in explaining Y. For example, in polynomial regression one usually would include a certain power of the explanatory variable only if all lower order terms are also included. If we assume, without loss of generality, that the ordering of the columns of the matrix X reflects the a priori given ordering, then this amounts to considering

$$\mathcal{M}_{nested} = \{M(p) : 0 \leq p \leq K\}$$

as the set of candidate models where

$$M(p) = \left\{(\theta, F) \in \mathbb{R}^K \times \mathcal{F} : \theta_i = 0 \text{ for } i > p\right\}.$$

Note that in this case the models $M(p)$ are nested in the sense that $M(p) \subseteq M(p+1)$ holds. Yet another variant is the set of candidate models

$$\mathcal{M}_{nested,protected} = \{M(p) : k \leq p \leq K\}$$

which obviously protects the first k variables in the context of nested model selection. If M is now a submodel of (1), one would typically estimate the parameters of model M by the (restricted) least squares estimator $\hat{\theta}(M)$ associated with M. Given a model selection procedure $\hat{M}$ selecting from a set $\mathcal{M}$ of candidate models, the associated PMSE is then given by

$$\tilde{\theta} = \sum_{M \in \mathcal{M}} \hat{\theta}(M)\mathbf{1}(\hat{M} = M). \tag{2}$$

For an extension of this example to the case of infinitely many regressors see Section 3.

Example 2. (Linear restrictions) Suppose the overall model is again given by model (1) but the submodels are now defined by general linear restrictions of the form $R\theta = r$.

Example 3. (Time Series Models) Suppose the data $Y = (y_1, \ldots, y_n)'$ follow an autoregressive model

$$y_t = \theta_1 y_{t-1} + \ldots + \theta_P y_{t-P} + u_t$$

for $t \geq 1$ and initial values $y_0, \ldots, y_{1-P}$. Typical assumptions on the errors u_t are that they are independent identically distributed according to a distribution with mean zero, or that the errors form a martingale difference sequence, etc. Of interest here are those submodels where $\theta_{p+1} = \theta_{p+2} = \ldots = \theta_P = 0$, in which case the model selection problem is the problem of selecting the order of the autoregressive process. [Similarly, the order selection problem for other classes of time series models such as, e.g., autoregressive moving average models or GARCH models obviously also fits into the framework outlined above.] In this example we have assumed that y_t is generated by a finite-order

autoregressive model. Often finite-order autoregressive models are fitted to a time series, e.g., for the purpose of prediction, even if the time series is not a finite-order autoregression. In this case the order of the approximating autoregressive model has to be determined from the data, leading again to a model selection problem that falls under the umbrella of the general framework formulated above.

Example 4. (General parametric models) Starting from an overall parametric model $\{\mathbb{P}_\eta : \eta \in H\}$, submodels M_g are defined by restrictions $g(\eta) = 0$, i.e., $M_g = \{\mathbb{P}_\eta : \eta \in H,\ g(\eta) = 0\}$, for g belonging to a given class $\mathcal{G}$ of restrictions. The set $\mathcal{M}$ is then given by $\mathcal{M} = \{M_g : g \in \mathcal{G}\}$. Note that models corresponding to different restrictions g will in general not be nested in each other, although they are nested in the overall model.

1.2 Model Selection Procedures

1.2.1 Procedures Based on Tests

Consider for simplicity first the case of only two competing candidate models M_i, $i = 1, 2$, where one is nested in the other, e.g., $M_1 \subseteq M_2$. Furthermore, assume that at least the larger model M_2 is correct, i.e., that the true probability distribution of Y belongs to M_2. Then a decision between the models M_1 and M_2 can be based on a test of the hypothesis H_0 that the true probability distribution belongs to M_1 versus the alternative H_1 that it belongs to $M_2 \backslash M_1$. More formally, let $\mathfrak{R}$ be a rejection region of a test for the hypothesis H_0. Then the selected model $\hat{M}$ is given by

$$\hat{M} = \begin{cases} M_1 \text{ if } Y \notin \mathfrak{R} \\ M_2 \text{ if } Y \in \mathfrak{R} \end{cases}.$$

For example, if M_2 corresponds to the linear model (1) with independent identically $N(0, \sigma^2)$ distributed errors and M_1 is given by a linear restriction $R\theta = r$, then the rejection region $\mathfrak{R}$ could be chosen as the rejection region of a classical F-test of this linear restriction.

In case of more than two candidate models which are nested, i.e., $M_1 \subseteq M_2 \subseteq \ldots \subseteq M_s$ holds, model selection can be based on a sequence of tests. For example, one can start by testing M_{s-1} against M_s. If this test rejects, one sets $\hat{M} = M_s$. Otherwise, a test of M_{s-2} against M_{s-1} is performed. In case this second test rejects, one sets $\hat{M} = M_{s-1}$. If this second test does not reject, one proceeds with testing M_{s-3} against M_{s-2} and so on, until a test rejects or one has reached the smallest model M_1. Such a procedure is often called a 'general-to-specific' procedure. Of course, one could also start from the smallest model and conduct a 'specific-to-general' testing procedure. If the set $\mathcal{M}$ of candidate models is not ordered by the inclusion relation ('non-nested case'), testing procedures can still be used to select a model $\hat{M}$ from $\mathcal{M}$, although then more thought has to be given to the order in which to conduct the tests

between competing models. The familiar stepwise regression procedures (see, e.g., Chapter 6 in Draper and Smith (1981) or Hocking (1976)) are a case in point. Model selection procedures based on hypothesis tests have been considered, e.g., in Anderson (1962, 1963), McKay (1977), Pötscher (1983, 1985), Bauer, Pötscher, and Hackl (1988), Hosoya (1984, 1986), and Vuong (1989); for a more recent contribution see Bunea, Wegkamp, and Auguste (2006). Also the related literature on pre-test estimators as summarized in Bancroft and Han (1977), Judge and Bock (1978), and Giles and Giles (1993) fits in here.

Returning to the case of two nested models M_1 and M_2, we note that the model selection procedures sketched above are based on testing whether the true distribution of Y belongs to model M_1 or not. However, if the goal is not so much selection of the 'true' model but is selection of a model that results in estimators with small mean squared error, it may be argued that the appropriate hypothesis to test is not the hypothesis that the distribution of Y belongs to M_1, but rather is the hypothesis that the mean squared error of the estimator based on M_1 is smaller than the mean squared error of the estimator based on M_2. Note that this is not the same as the hypothesis that the distribution of Y belongs to M_1. This observation seems to have first been made by Toro-Vizcarrondo and Wallace (1968), see also Wallace (1972). In the context where M_2 is a normal linear regression model and M_1 is given by a linear restriction $R\theta = r$, they show that the mean squared error matrix of the restricted least squares estimator is less than or equal to the mean squared error matrix of the unrestricted least squares estimator whenever $\sigma^{-2}\theta' R'[R(X'X)^{-1}R']^{-1}R\theta \leq 1$ holds. Hence, they propose to select the model M_1 whenever a test for the hypothesis $\sigma^{-2}\theta' R'[R(X'X)^{-1}R']^{-1}R\theta \leq 1$ does not reject, and to select M_2 otherwise. It turns out that the appropriate test statistic is again the F-statistic, but with a critical value that is chosen from a non-central F-distribution.

It is important to point out that the PMSE (aka 'pre-test' estimator) for θ resulting from first selecting the model $\hat{M}$ by some of the testing procedures described above and then estimating the parameters in the model $\hat{M}$ by least squares is neither the restricted nor the unrestricted least squares estimator, but a *random* convex combination of both, cf. (2). In particular, while it is true that the mean squared error of the restricted least squares estimator (corresponding to M_1) is smaller than the mean squared error of the unrestricted least squares estimator (corresponding to M_2) whenever model M_1 is true (and more generally, as long as $\sigma^{-2}\theta' R'[R(X'X)^{-1}R']^{-1}R\theta \leq 1$ holds), the PMSE need not (and will not) have a mean squared error equal to the better of the mean squared errors of the restricted and unrestricted estimators, but will be larger. Hence, if keeping mean squared error of the PMSE small is the ultimate goal, one should set the significance level for the test underlying the model selection procedure such that the overshoot over the better of the mean squared errors of the restricted and unrestricted estimators does not exceed a prescribed 'tolerance level'. This has been investigated by Kennedy and Ban-

croft (1971), Sawa and Hiromatsu (1973), Brook (1976), Toyoda and Wallace (1976), Droge (1993), and Droge and Georg (1995). See also Amemiya (1980, Section 10).

1.2.2 Procedures Based on Model Selection Criteria

If the ultimate goal of model selection is to find a model that gives rise to parameter estimators or predictors with small mean squared error (or some other risk measure) it seems to be natural to approach the model selection problem in a way that is geared towards this aim. The approach of Toro-Vizcarrondo and Wallace (1968) mentioned above combines the testing approach with such a risk-oriented approach. Alternatively, one can try to estimate the model associated with the smallest risk. [Whether or not the ensuing PMSE then actually has small risk is another matter, see the discussion further below.] To fix ideas consider the standard linear regression model (1) with errors that have mean zero and variance-covariance matrix $\sigma^2 I_n$. For any model $M \in \mathcal{M}_{all}$ let $\hat{\theta}(M)$ denote the (restricted) least squares estimator computed under the zero-restrictions defining M. The mean squared error of $X\hat{\theta}(M)$ is then given by

$$\begin{aligned} MSE_{n,\theta}(M) &= \mathbb{E}_{n,\theta} \left\| X\hat{\theta}(M) - X\theta \right\|^2 = \mathbb{E}_{n,\theta} \left\| P_M Y - X\theta \right\|^2 \\ &= \sigma^2 \operatorname{tr}(P_M) + \theta' X'(I - P_M) X\theta \\ &= \sigma^2 k_M + \theta' X'(I - P_M) X\theta \end{aligned} \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm, P_M denotes projection on the column space spanned by the regressors active in M, and k_M denotes the number of these regressors. Ideally, we would like to use that model M that minimizes the risk (3), i.e., the model that has mean squared error equal to

$$\min_{M \in \mathcal{M}} MSE_{n,\theta}(M) \tag{4}$$

where $\mathcal{M}$ is the set of candidate models specified by the researcher. The expression in (4) is sometimes called the 'risk-target' and it depends on the unknown parameters θ and σ^2 as well as on the set of candidate models $\mathcal{M}$ (and on X). However, since (3) (and (4)) are unobservable, it is not feasible to use the risk-minimizing model. An obvious idea is then to estimate (3) for every $M \in \mathcal{M}$ and to find the model that minimizes this estimator of the risk (sometimes called the 'empirical risk'). An unbiased estimator for (3) is easily found as follows: Let M_{full} denote model (1), i.e., the model containing all K regressors and let $\hat{\theta}$ be shorthand for $\hat{\theta}(M_{full})$. Then

$$\begin{aligned} \mathbb{E}_{n,\theta} \left(\hat{\theta}' X'(I - P_M) X\hat{\theta} \right) &= \mathbb{E}_{n,\theta} \left(Y' P_{M_{full}} (I - P_M) P_{M_{full}} Y \right) \\ &= \mathbb{E}_{n,\theta} \left(Y' (P_{M_{full}} - P_M) Y \right) \\ &= \sigma^2 (K - k_M) + \theta' X'(I - P_M) X\theta. \end{aligned}$$

Since σ^2 can easily be estimated unbiasedly by $\hat{\sigma}^2 = \hat{\sigma}^2(M_{full}) = (n - K)^{-1}Y'(I - P_{M_{full}})Y$, an unbiased estimator for $MSE_{n,\theta}(M)$ is found to be

$$\mathsf{MC}_n(M) = \hat{\theta}' X'(I - P_M)X\hat{\theta} + 2k_M\hat{\sigma}^2 - K\hat{\sigma}^2. \tag{5}$$

Noting that $X\hat{\theta}$ equals $P_{M_{full}}Y$ this can be rewritten as

$$\mathsf{MC}_n(M) = RSS(M) + 2k_M\hat{\sigma}^2 - n\hat{\sigma}^2 \tag{6}$$

where $RSS(M) = Y'(I - P_M)Y$. After division by $\hat{\sigma}^2$, this is known as Mallows' C_p introduced in 1964; see Mallows (1965, 1973). The model selection procedure based on Mallows' C_p now returns that model $\hat{M}$ which minimizes (6) over the set $\mathcal{M}$. It should be mentioned that Mallows did not advocate the minimum C_p strategy just described, but voiced concern about this use of C_p (Mallows (1965, 1973, 1995)).

It is important to note that the PMSE $\tilde{\theta}$ for θ obtained via selection of the model minimizing (6) is a compound procedure, and is *not* identical to any of the least squares estimators $\hat{\theta}(M)$ obtained from the models $M \in \mathcal{M}$; as pointed out before in (2), it rather is a *random* convex combination of these estimators. As a consequence, despite the construction of $\hat{M}$ as a minimizer of an empirical version of the risk of the least squares estimators associated with the models M, it does *not* follow that the mean squared error of $\tilde{\theta}$ is equal to (or close to) the risk target (4). In fact, it can overshoot the risk target considerably. This comment applies mutatis mutandis also to the model selection procedures discussed below and we shall return to this issue also later on in Section 2.2.

A related criterion is the so-called 'final prediction error' (FPE) which has become well-known through the work of Akaike (1969, 1970), set in the context of selecting the order of autoregressive models. The same criterion was actually introduced earlier by Davisson (1965) also in a time series context, and was – according to Hocking (1976) – discussed by Mallows (1967) in a regression context. In the present context of a linear regression model it amounts to selecting the model M that minimizes

$$\mathsf{FPE}_n(M) = RSS(M)(n - k_M)^{-1}(1 + k_M/n). \tag{7}$$

The derivation of FPE is somewhat similar in spirit to the derivation of Mallows' C_p: Suppose that now the mean squared error of prediction

$$\begin{aligned} MSEP_{n,\theta}(M) &= \mathbb{E}_{n,\theta}\left\|Y^* - X\hat{\theta}(M)\right\|^2 \\ &= \sigma^2(n + k_M) + \theta' X'(I - P_M)X\theta \end{aligned} \tag{8}$$

is the quantity of interest, where $Y^* = X\theta + u^*$ with u^* having the same distribution as u, but is independent of u. [Note that in the linear regression model considered here $MSE_{n,\theta}(M)$ and $MSEP_{n,\theta}(M)$ only differ by the additive term $\sigma^2 n$, hence this difference is immaterial and we have switched to

$MSEP_{n,\theta}(M)$ only to be in line with the literature.] For models M that are correct, the second term in (8) vanishes and – transposing Akaike's (1970) argument in the autoregressive case to the case of linear regression – it is proposed to estimate the unknown variance σ^2 in the first term by $\hat{\sigma}^2(M) = (n - k_M)^{-1} RSS(M)$. Upon division by n, this gives (7). Hence, $n\mathsf{FPE}_n(M)$ is an unbiased estimator for (8) *provided* the model M is correct. For incorrect models M this is not necessarily so, but it is suggested in Akaike (1969, 1970) that then the misspecification bias will make $\hat{\sigma}^2(M)$ large, obviating the need to take care of the bias term $\theta' X'(I - P_M)X\theta$. While this is true for fixed θ and large n, ignoring the bias term seems to be an unsatisfactory aspect of the derivation of FPE. [Note also that if one would estimate σ^2 by $\hat{\sigma}^2 = \hat{\sigma}^2(M_{full})$ rather than $\hat{\sigma}^2(M)$ in the above derivation, one would end up with the absurd criterion $\hat{\sigma}^2(1 + k_M/n)$.]

Akaike's (1973) model selection criterion AIC is derived by similar means and – in contrast to Mallows' C_p or FPE, which are limited to linear (auto)regressions – is applicable in general parametric models. The risk measure used here is not mean squared error of prediction but the expected Kullback-Leibler discrepancy between $\mathbb{P}_{\hat{\eta}(M)}$ and the true distribution of Y, where $\hat{\eta}(M)$ denotes the maximum likelihood estimator based on model M. Akaike (1973) proposes an estimator for the Kullback-Leibler discrepancy that is approximately unbiased *provided* that the model M is a correct model. This estimator is given by $(n/2)\mathsf{AIC}_n(M)$ where

$$\mathsf{AIC}_n(M) = -2n^{-1} \log L_{n,M}(Y, \hat{\eta}(M)) + 2\#M/n, \tag{9}$$

$L_{n,M}$ denotes the likelihood function corresponding to model M, and $\#M$ denotes the number of parameters in M. [The analysis in Akaike (1973) is restricted to i.i.d. data, but can be extended to more general settings; see, e.g., Findley (1985) for a treatment in the context of linear time series models, and Findley and Wei (2002) for vector autoregressive models.] The minimum AIC-procedure now consists of selecting that model $\hat{M}$ that minimizes AIC_n over the set $\mathcal{M}$. For the linear regression model (1) with errors $u \sim N(0, \sigma^2 I_n)$, σ^2 unknown, the criterion AIC_n reduces – up to an irrelevant additive constant – to

$$\mathsf{AIC}_n(M) = \log(RSS(M)/n) + 2k_M/n. \tag{10}$$

[If the error variance σ^2 is known, $\mathsf{AIC}_n(M)$ is – again up to an irrelevant additive constant – equal to $\mathsf{MC}_n(M)$ with $\hat{\sigma}^2$ replaced by σ^2.] For a very readable account of the derivation of the criteria discussed so far see Amemiya (1980).

A different approach to model selection, which is Bayesian in nature, was taken by Schwarz (1978). Given priors on the parameters in each model M and prior probabilities for each model (i.e., a prior on $\mathcal{M}$) one can compute the posterior probability for each model M given the data and one would then choose the model with the highest posterior probability. Schwarz (1978) showed that the leading terms in the posterior probabilities do not depend on

the specific prior employed: He showed that the negative of the log posterior probabilities can – for large sample sizes – be approximated by $(n/2)\mathsf{BIC}_n(M)$ where

$$\mathsf{BIC}_n(M) = -2n^{-1} \log L_{n,M}(Y, \hat{\eta}(M)) + \#M(\log n)/n. \tag{11}$$

The minimum BIC procedure then selects the model $\hat{M}$ that minimizes $\mathsf{BIC}_n(M)$ over $\mathcal{M}$.

Variants of the procedures. A variant of FPE, studied in Bhansali and Downham (1977), is FPE_α which reduces to FPE for $\alpha = 2$, see also Shibata (1984). Shibata (1986b) and Venter and Steele (1992) discuss ways of choosing α such that the maximal (regret) risk of the ensuing PMSE is controlled; cf. also Foster and George (1994). Variants of AIC/BIC obtained by replacing the $\log n$ term in (11) by some other function of sample size have been studied, e.g., in Hannan and Quinn (1979), Pötscher (1989), Rao and Wu (1989), and Shao (1997); cf. also Section 2.1. As noted, the AIC criterion is an asymptotically unbiased estimator of the Kullback-Leibler discrepancy if the model M is correct. A finite-sample bias correction for correct models M has been provided by Sugiura (1978) and subsequently by Hurvich and Tsai (1989) and leads to the so-called AICC criterion which in the Gaussian linear regression context takes the form

$$\mathsf{AICC}_n(M) = \log(RSS(M)/n) + 2(k_M + 1)/(n - k_M - 2).$$

[The derivation of AIC or AICC from asymptotically unbiased estimators for the Kullback-Leibler discrepancy is based on the assumption that the models M are correct models. For bias corrections allowing for M to be incorrect and for resulting model selection criteria see Reschenhofer (1999) and references therein.] Takeuchi's (1976) criterion TIC should also be mentioned here which is an approximately unbiased estimator of Kullback-Leibler discrepancy also for incorrect models. However, TIC requires consistent estimators for the expectations of the Hessian of the log-likelihood as well as of the outer product of the score, where the expectation is taken under the true distribution. A possibility to implement this is to use bootstrap methods, see Shibata (1997). The derivations underlying AIC or TIC assume that the models are estimated by maximum likelihood. Konishi and Kitagawa (1996) introduced a model selection criterion GIC that allows for estimation procedures other than maximum likelihood. See also the recent book by Konishi and Kitagawa (2008). The derivation of FPE in autoregressive models is based on the one-step-ahead prediction error. Model selection criteria that focus on multi-step-ahead predictors can similarly be derived and are treated in Findley (1991), Bhansali (1999), and Ing (2004).

Other model selection criteria. Myriads of model selection criteria have been proposed in the literature and it is impossible to review them all. Here we just want to mention some of the more prominent criteria not yet discussed. Theil (1961) was perhaps the first to suggest to use the adjusted R^2 as a model selection criterion. Maximization of the adjusted R^2 amounts to minimization

(w.r.t. M) of

$$(n - k_M)^{-1} RSS(M).$$

Another early criterion that has apparently been suggested by Tukey is given by

$$S_n(M) = ((n - k_M)(n - k_M - 1))^{-1} RSS(M). \tag{12}$$

It is – similar to Mallows' C_p – obtained from an unbiased estimator of the out-of-sample mean squared error of prediction in a linear regression model, where now the vector of regressors is assumed to be independent and identical normally distributed with mean zero and the expectation defining the mean squared error of prediction is also taken over the regressors (in the observation as well as in the prediction period). This criterion is further discussed in Thompson (1978a,b), Breiman and Freedman (1983), and Leeb (2006b). Cross-validation provides another method for model selection (Allen (1974), Stone (1974), Shao (1993), Zhang (1993a), Rao and Wu (2005)). Generalized cross-validation has been introduced by Craven and Wahba (1979) and in the linear regression context of Example 1 amounts to minimizing

$$GCV_n(M) = (n - k_M)^{-2} RSS(M).$$

Note the close relationship with Tukey's $S_n(M)$. Also in the context of a standard linear regression model with non-stochastic regressors, Foster and George (1994) introduced the so-called risk inflation criterion based on considering the minimization of the maximal inflation of the risk of the PMSE over the infeasible 'estimator' that makes use of the knowledge of the (minimal) true model. This criterion is given by

$$\mathsf{RIC}_n(M) = RSS(M) + 2k_M \log(K)\hat{\sigma}^2.$$

See also George and Foster (2000). Based on considerations of the code length necessary to encode the given data by encoding the fitted candidate models, Rissanen (1978, 1983, 1986a,b, 1987) has introduced the minimum description length (MDL) criterion and the closely related predictive least squares (PLS) criterion; cf. also the review article by Hansen and Yu (2001) as well as Rissanen (1989). These criteria are also closely connected to Wei's (1992) Fisher information criterion (FIC). [For more on this criterion see the article by N. Chan in this *Handbook*.] Finally, if prediction at a value $x_{f\cdot}$ of the regressor vector different from the values in the sample is of interest and if the steps in the derivation of Mallows' C_p are repeated for this target, one ends up with a criterion introduced in Allen (1971). This criterion depends on the chosen $x_{f\cdot}$ and hence is an early precursor to the so-called focused information criterion of Claeskens and Hjort (2003). For a discussion of further model selection criteria see Rao and Wu (2001).

Relationships between criteria. For many model selection problems such as, e.g., variable selection in linear regression or order selection for autoregressive processes the criteria AIC, AICC, FPE, Mallows' C_p, Tukey's S_n, cross-validation as well as generalized cross-validation are 'asymptotically equivalent' (Stone (1977), Shibata (1989)). These asymptotic equivalence results

typically hold only for quite 'small' families $\mathcal{M}$ of candidate models (e.g., for fixed finite families); in particular, k_M typically has to be small compared to n for the asymptotic equivalence results to bear on the finite-sample behavior. If k_M is not small relative to n, the asymptotic equivalence does not apply and these criteria can behave very differently. For more discussion on the relationship between various criteria see, e.g., Söderström (1977), Amemiya (1980), Teräsvirta and Mellin (1986), and Leeb (2006b).

A Comment. Criteria like Mallows' C_p, AIC, FPE, and so on have been derived as (asymptotically) unbiased estimators for mean squared error (of prediction) or Kullback-Leibler discrepancy for certain estimation problems. Often these criteria are also used in contexts where they are not necessarily (asymptotically) unbiased estimators (e.g., in a pseudo-likelihood context), or where no formal proof for the unbiasdness property has been provided. For example, for Gaussian AR models AIC is (approximately) unbiased (Findley and Wei (2002)) and takes the form $\log \hat{\sigma}^2(k) + 2k/n$ where $\hat{\sigma}^2(k)$ is the usual residual variance estimator from an AR(k)-fit. This latter formula is, however, routinely also used for model selection in AR models with non-Gaussian (even heavy-tailed) errors without further justification. Another example is model selection in GARCH models, where procedures like minimum AIC are routinely applied, but formal justifications do not seem to be available.

Relationship between model selection criteria and hypothesis tests. The model selection procedures described in Section 1.2.1 and in the present section are closely related. First observe that in a setting with only two nested models, i.e., $M_1 \subseteq M_2$, the minimum AIC-procedure picks model M_2 if and only if the usual likelihood ratio test statistic of the hypothesis M_1 versus M_2 exceeds the critical value $2(k_{M_2} - k_{M_1})$. In general, the model M selected by minimum AIC is characterized by the property that the likelihood ratio test statistic for testing M against any other model $M' \in \mathcal{M}$ (nesting M or not) exceeds the respective critical value $2(k_{M'} - k_M)$. For more discussion see Söderström (1977), Amemiya (1980), Teräsvirta and Mellin (1986), and Pötscher (1991, Section 4).

2 Properties of Model Selection Procedures and of Post-Model-Selection Estimators

We now turn to the statistical properties of model selection procedures and their associated PMSEs. In particular, questions like consistency/inconsistency of the model selection procedure, risk properties, as well as distributional properties of the associated PMSE are discussed. In this section we concentrate on the case where the set $\mathcal{M}$ of candidate models contains a correct model; furthermore, $\mathcal{M}$ is here typically assumed to be finite, although some of the results mentioned in this section also hold if $\mathcal{M}$ expands suitably with sample size or is infinite. The case of model selection from a set of models that are po-

tentially only approximations to the data generating mechanism is discussed in Section 3.

2.1 Selection Probabilities and Consistency

The focus in this subsection is on the model selection procedure $\hat{M}$ viewed as an estimator for the minimal true model (given that it exists). For definiteness of discussion, consider the linear regression model as in Example 1 with an $N(0, \sigma^2 I_n)$-distributed error term. Assume for simplicity of presentation further that the set $\mathcal{M} \subseteq \mathcal{M}_{all}$ of candidate models contains the full model M_{full} and is stable w.r.t. intersections, meaning that with M and M' belonging to $\mathcal{M}$, also $M \cap M'$ belongs to $\mathcal{M}$. [This is, e.g., the case for $\mathcal{M} = \mathcal{M}_{all}$ or $\mathcal{M} = \mathcal{M}_{nested}$.] Under this condition, for each value of the parameter $\theta \in \mathbb{R}^K$ there exists a minimal true model $M_0 = M_0(\theta)$ given by

$$M_0 = \bigcap_{\theta \in M \in \mathcal{M}} M.$$

[If $\mathcal{M} = \mathcal{M}_{all}$, then M_0 is given by the set of all parameters θ^* that have $\theta_i^* = 0$ whenever $\theta_i = 0$. If $\mathcal{M} = \mathcal{M}_{nested}$, then M_0 is given by the set of all parameters θ^* that have $\theta_i^* = 0$ for all $i > p_0(\theta)$, where $p_0(\theta)$ is the largest index such that $\theta_{p_0(\theta)} \neq 0$ (and $p_0(\theta) = 0$ if $\theta = 0$).] The quality of $\hat{M}$ as an estimator for M_0 can be judged in terms of the 'overestimation' and 'underestimation' probabilities, respectively, i.e., in terms of the probabilities of the events

$$\{\hat{M} \neq M_0, \hat{M} \supseteq M_0\} \tag{13}$$

and

$$\{\hat{M} \not\supseteq M_0\}. \tag{14}$$

Note that (13) represents the case where a correct model containing superfluous regressors is selected, whereas (14) describes the case where an incorrect model is selected.

A model selection procedure is *consistent* if the probabilities of over- and underestimation converge to zero, i.e., if

$$\lim_{n \to \infty} P_{n,\theta}(\hat{M} = M_0) = 1 \tag{15}$$

for every $\theta \in \mathbb{R}^K$. If

$$\lim_{n \to \infty} P_{n,\theta}(\hat{M} \not\supseteq M_0) = 0 \tag{16}$$

for every $\theta \in \mathbb{R}^K$, but $\hat{M}$ is not consistent, we say that $\hat{M}$ is *conservative*. We note that any reasonable model selection procedure will satisfy (16). Suppose that in the context of the linear regression model considered here the regressors also satisfy the 'asymptotic stationarity' condition $X'X/n \to Q$, where Q is a positive definite matrix. If $\hat{M}$ is then obtained through minimization of a criterion of the form

$$\mathsf{CRIT}(M) = \log(RSS(M)/n) + k_M C_n/n, \tag{17}$$

it is well-known that $\hat{M}$ is consistent if the penalty satisfies $C_n/n \to 0$ and $C_n \to \infty$ as $n \to \infty$; and it is conservative if C_n is bounded (e.g., Geweke and Meese (1981)). In particular, it follows that the minimum BIC procedure (i.e., $C_n = \log n$) is consistent, whereas the minimum AIC procedure (i.e., $C_n = 2$) is conservative. [That FPE is conservative has already been noted in Akaike (1970) in the context of selecting the order of stationary autoregressions.] If the asymptotic stationarity condition $X'X/n \to Q$ does not hold, the conditions on C_n for consistency/conservatism change and are related to the rate of increase of the largest and smallest eigenvalues of $X'X$ (Pötscher (1989)). We note that these results are not tied to the normality or i.i.d. assumption on the errors made here for the sake of simplicity, but hold under more general assumptions. For further consistency results in the context of linear regression models see Nishii (1984), An and Gu (1985), Rao and Wu (1989), and Zheng and Loh (1995, 1997); Chen and Ni (1989) provide consistency results for linear regression models with short-memory time series errors, while Ing and Wei (2006) allow also for long-memory errors; see also Hidalgo (2002). Consistency results for model selection based on (17) or on closely related criteria, but for model classes other than linear regression, can be found in Hannan and Quinn (1979) and Quinn (1980) for stationary autoregressions and in Hannan (1980, 1981) for stationary autoregressive moving average (ARMA) models (see also An and Chen (1986) and Chapter 5 of Hannan and Deistler (1988) for more discussion and references); in Paulsen (1984), Tsay (1984), Pötscher (1989), and Wei (1992) for nonstationary autoregressions, the latter two papers considering also more general classes of stochastic regression models; in Knight (1989) for infinite variance autoregressions; in Kohn (1983) and Nishii (1988) for general parametric models; and in Haughton (1991) for nonlinear regression models. Similar results hold for criteria like FPE_α if α is made dependent on sample size in an appropriate manner and are discussed in several of the references just given. Consistency results for the PLS criterion can be found, e.g., in Rissanen (1986b), Hemerly and Davis (1989), and Wei (1992). The papers on consistency mentioned so far consider a finite set of candidate models (which in some results is allowed to expand with sample size, typically slowly). In the context of order selection of stationary ARMA models, Pötscher (1990) discusses a modification of BIC-like procedures and establishes consistency without any restriction on the size of the set of candidate models, i.e., the result applies even for $\mathcal{M}$ the (infinite) set of *all* ARMA models; see also Pötscher and Srinivasan (1994). We are not aware of any published formal results establishing consistency of model selection procedures in GARCH models, although such results are certainly possible. Francq, Roussignol, and Zakoïan (2001) establish that AIC and related criteria are conservative procedures in the context of ARCH models.

Model selection procedures based on tests are typically consistent if the critical values employed by the tests are chosen in such a way that they diverge

to infinity at an appropriate rate. Otherwise the procedures are typically conservative. Consistency results of this sort are provided in Pötscher (1983) in the context of selecting the order of ARMA models and by Bauer, Pötscher, and Hackl (1988) for a 'thresholding' procedure in a general (semi)parametric model. For a follow-up on the latter paper see Bunea, Wegkamp, and Auguste (2006).

The limits of the model selection probabilities in case of conservative model selection procedures have been studied in Shibata (1976), Bhansali and Downham (1977), Hannan (1980), Geweke and Meese (1981), Sakai (1981), Tsay (1984), and Quinn (1988). Further studies of the over- and underestimation probabilities in various settings and for various model selection procedures can be found in Zhang (1993b), Shao (1998), Guyon and Yao (1999), and Keribin and Haughton (2003).

Since it seems to be rarely the case that estimation of M_0 is the ultimate goal of the analysis, the consistency property of $\hat{M}$ may not be overly important. In fact, as we shall see in the next subsection, consistency of $\hat{M}$ has detrimental effects on the risk properties of the associated PMSE. This may seem to be counterintuitive at first sight and is related to the fact that the consistency property (15) does not hold uniformly w.r.t. the parameter θ (Leeb and Pötscher (2005), Remark 4.4). Furthermore, in a situation where none of the models in the class $\mathcal{M}$ is correct (see Section 3), that is if only 'approximate' models are fitted, the concept of consistency becomes irrelevant.

2.2 Risk Properties of Post-Model-Selection Estimators

As already noted in Sections 1.2.1 and 1.2.2 it is important to realize that – despite the ideas underlying the construction of PMSEs – a PMSE does not come with an automatic optimality property; in particular, its risk is by no means guaranteed to equal the risk target (4). For example, while $\mathsf{MC}_n(M)$ given by (5) is an unbiased estimator of the mean squared error $MSE_{n,\theta}(M)$ of the estimator $\hat{\theta}(M)$ based on model M, minimizing $\mathsf{MC}_n(M)$ gives rise to a *random* $\hat{M}$ and an associated PMSE $\tilde{\theta}$ that is a *random* convex combination of the estimators $\hat{\theta}(M)$ based on the various models $M \in \mathcal{M}$. As a consequence, the mean squared error of $\tilde{\theta}$ is no longer described by any of the quantities $MSE_{n,\theta}(M)$ (or by the risk target (4) for that matter), since $\tilde{\theta}$ falls outside of the class $\{\hat{\theta}(M) : M \in \mathcal{M}\}$.

The risk-properties of PMSEs have been studied for model selection procedures based on test procedures in considerable detail, see Judge and Bock (1978), Giles and Giles (1993), Magnus (1999), and Danilov and Magnus (2004). For procedures based on model selection criteria investigations into the risk-properties of PMSEs can be found in Mallows (1973), Hosoya (1984), Nishii (1984), Shibata (1984, 1986b, 1989), Venter and Steele (1992), and Foster and George (1994). A basic feature of risk-functions of PMSEs that emerges from these studies is best understood in the context of the simple normal linear regression model (Example 1) when selection is only between

two models M_1 and $M_2 = M_{full}$ where M_1 is obtained from M_2 by restricting the $k_2 \times 1$ subvector θ_2 of the $(k_1 + k_2) \times 1$ parameter vector $\theta = (\theta_1', \theta_2')'$ to zero. Recall that in this example the quadratic risk $MSE_{n,\theta}(\hat{\theta}(M_1))$ is an unbounded quadratic function of θ_2, which achieves its minimal value $\sigma^2 k_1$ on the set $\{\theta : \theta_2 = 0\}$, i.e., when model M_1 holds, whereas the quadratic risk $MSE_{n,\theta}(\hat{\theta}(M_2))$ is constant and equals $\sigma^2(k_1 + k_2)$. Hence, $MSE_{n,\theta}(\hat{\theta}(M_1)) < MSE_{n,\theta}(\hat{\theta}(M_2))$ whenever $\theta_2 = 0$, and this inequality persists for $\theta_2 \neq 0$ sufficiently small (by continuity of the mean squared error). However, as θ_2 moves away from the origin, eventually the inequality will be reversed. Now, the risk $MSE_{n,\theta}(\tilde{\theta})$ of the PMSE $\tilde{\theta}$ will typically also be less than the risk of $\hat{\theta}(M_2)$ for parameter values which have $\|\theta_2\|$ sufficiently close to zero, but it will rise *above* the risk of $\hat{\theta}(M_2)$ as $\|\theta_2\|$ becomes larger; eventually the risk of $\tilde{\theta}$ will attain its maximum and then gradually approach the risk of $\hat{\theta}(M_2)$ from above as $\|\theta_2\|$ increases further and approaches infinity. As stressed by Magnus (1999), for many PMSEs there will be even regions in the parameter space (for intermediate values of $\|\theta_2\|$) where the risk of the PMSE will actually be larger than the larger of the risks of $\hat{\theta}(M_1)$ and $\hat{\theta}(M_2)$. Figure 5 in Leeb and Pötscher (2005) gives a representation of the typical risk behavior of a PMSE.

Limiting risk in case of consistent model selection. Continuing the above example, suppose $\hat{M}$ is a consistent model selection procedure for M_0, i.e., $\hat{M}$ satisfies (15), where the minimal true model M_0 is here given by

$$M_0 = \begin{cases} M_1 \text{ if } \theta_2 = 0 \\ M_2 \text{ if } \theta_2 \neq 0 \end{cases}.$$

Assume also that $X'X/n \to Q > 0$ for $n \to \infty$. Then $P_{n,\theta}(\hat{M} = M_1)$ will typically go to zero exponentially fast for $\theta_2 \neq 0$ (Nishii (1984)). It is then easy to see that

$$\lim_{n\to\infty} MSE_{n,\theta}(\tilde{\theta}) = \begin{cases} \sigma^2 k_1 & \text{if } \theta_2 = 0 \\ \sigma^2(k_1 + k_2) & \text{if } \theta_2 \neq 0 \end{cases}. \tag{18}$$

That is, for each *fixed* θ the limiting risk of the PMSE coincides with the (limiting) risk of the restricted estimator $\hat{\theta}(M_1)$ if model M_1 obtains, and with the (limiting) risk of the unrestricted estimator $\hat{\theta}(M_2)$ if model M_2 is the minimal true model. Put yet another way, the limiting risk of the PMSE coincides with the (limiting) risk of the 'oracle', i.e., of the infeasible 'estimator' based on the minimal true model M_0. This *seems* to tell us that in large samples consistent model selection typically results in a PMSE that has approximately the same risk behavior as the *infeasible* procedure that uses $\hat{\theta}(M_1)$ if $\theta_2 = 0$ and uses $\hat{\theta}(M_2)$ if $\theta_2 \neq 0$. Note that this procedure is infeasible (and hence is sometimes dubbed an 'oracle'), since it uses knowledge of whether $\theta_2 = 0$ or not. The above observation that in a 'pointwise' asymptotic analysis (that is in an asymptotic analysis that holds the true parameter fixed while letting sample

size increase) a consistent model selection procedure typically has no effect on the limiting risk of the PMSE has been made in Nishii (1984); see also the discussion of the so-called 'oracle property' in Section 2.3. Unfortunately, this result – while mathematically correct – is a statistical fallacy and does not even approximately reflect the risk properties at any given sample size, regardless how large: It can be shown that – despite (18) – the worst-case risk of *any* PMSE based on a consistent model selection procedure diverges to infinity, i.e.,

$$\lim_{n\to\infty} \sup_{\theta} MSE_{n,\theta}(\tilde{\theta}) = \infty \tag{19}$$

holds. Hence, in terms of worst-case risk a PMSE based on a consistent model selection procedure is much worse than, e.g., the least squares estimator based on the overall model (which has constant risk $\sigma^2(k_1 + k_2)$), or than a PMSE based on a conservative procedure (which typically has bounded worst-case risk, cf. (20) below). This phenomenon, which is in striking contrast to (18), has been observed at different levels of generality by Hosoya (1984), Shibata (1986b), Foster and George (1994), Leeb and Pötscher (2005, 2008a), and Yang (2005, 2007). As shown in Leeb and Pötscher (2008a), the unboundedness phenomenon (19) also applies to so-called 'sparse' estimators as considered, e.g., in Fan and Li (2001); cf. Section 4. We note that the finite-sample behavior of the risk function of the PMSE, which gets lost in a pointwise asymptotic analysis, can be captured in an asymptotic analysis the makes the true parameter dependent on sample size; see Leeb and Pötscher (2005).

Limiting risk in case of conservative model selection. Results on the limiting risk in this case have been obtained in Hosoya (1984), Nishii (1984), Shibata (1984), and Zhang (1992). For conservative model selection procedures the limiting risk of the associated PMSE does not satisfy (18). In fact, in this case it can be shown that the limiting risk of a PMSE is typically larger than the limiting risk of the corresponding oracle (i.e., the infeasible 'estimator' based on the minimal true model M_0), except if the minimal true model is the overall model. Hence, for conservative model selection procedures a pointwise asymptotic analysis already reveals some of the effects of model selection on the risk of the PMSE, although the full effect is again only seen in an asymptotic analysis that makes the true parameter dependent on sample size (or in a finite-sample analysis, of course); see Leeb and Pötscher (2005) for an extensive discussion. In contrast to PMSEs based on consistent model selection procedures, the worst-case risk of a PMSE based on a conservative procedure is typically bounded as sample size goes to infinity, i.e.,

$$\lim_{n\to\infty} \sup_{\theta} MSE_{n,\theta}(\tilde{\theta}) < \infty \tag{20}$$

typically holds.

Admissibility results. Admissibility or inadmissibility of PMSEs in various classes of estimators has been discussed in Sclove, Morris, and Radhakrishnan (1972), Stone (1981, 1982), Takada (1982), and Kempthorne (1984).

Consistency of PMSEs. A PMSE is typically consistent regardless of whether the model selection procedure is consistent or conservative; this follows, e.g., from Lemma 2 in Pötscher (1991). In fact, PMSEs will often be even uniformly consistent, cf. Leeb and Pötscher (2005), Propositions A.9 and B.1, and Pötscher and Leeb (2007), Theorem 2.

2.3 Distributional Properties of Post-Model-Selection Estimators

As noted in Section 1.1, a PMSE $\tilde{\eta}$ is a *random* convex combination of the estimators $\hat{\eta}(M)$ computed on the basis of model M. As a consequence, the distribution of $\tilde{\eta}$ will typically be more complex than the distribution of $\hat{\eta}(M)$, which often will be asymptotically normal. Sen (1979) derived the asymptotic distribution of $n^{1/2}(\tilde{\eta} - \eta)$ in a maximum likelihood framework for i.i.d. data when the model selection procedure consists in choosing from two nested models $M_1 \subseteq M_2$ on the basis of the likelihood ratio test. Pötscher (1991) obtained the asymptotic distribution for the case when model selection is from a nested family $M_1 \subseteq M_2 \subseteq \ldots \subseteq M_P$ and is based on a general-to-specific hypothesis testing scheme; the framework in Pötscher (1991) is also more general than the one in Sen (1979) in that it allows for dependent data and M-estimators other than maximum likelihood. Furthermore, Pötscher (1991) derived not only the unconditional, but also the conditional asymptotic distribution of $n^{1/2}(\tilde{\eta} - \eta)$. Here the conditioning is on the event of having chosen a particular model. See Pötscher and Novak (1998) for further results and a simulation study. In the same framework as in Pötscher (1991), but confining attention to the normal linear regression model, Leeb and Pötscher (2003) and Leeb (2005, 2006a) obtained the unconditional as well as conditional finite-sample distribution of $n^{1/2}(\tilde{\eta} - \eta)$ (as well as their limits under local alternatives). From the above references it transpires that the asymptotic as well as the finite-sample distributions of $n^{1/2}(\tilde{\eta} - \eta)$ are complicated and, in particular, are typically decidedly non-normal, e.g., they can be bimodal. Furthermore, these distributions depend on the unknown parameter η in a complicated way. As a consequence of these results, the usual confidence intervals naively applied to PMSEs do not have correct coverage probability, not even asymptotically (Saleh and Sen (1983), Pötscher (1991), Zhang (1992), Kabaila (1998), Kabaila and Leeb (2006)). For further results on distributional properties of PMSEs based on conservative model selection procedures see Sen and Saleh (1987), Dijkstra and Veldkamp (1988), Kabaila (1995), Pötscher (1995), Ahmed and Basu (2000), and Hjort and Claeskens (2003). Shen, Huang, and Ye (2004, Theorem 3) incorrectly claim that the asymptotic distribution of the PMSE based on, say, AIC, is normal.

The results discussed in this subsection so far apply to *conservative* model selection procedures. It is important to note, however, that the finite-sample results in Leeb and Pötscher (2003) and Leeb (2005, 2006a) also apply to *consistent* model selection procedures based on general-to-specific testing (i.e., procedures where the critical values diverge to infinity at an appropriate rate

with sample size), since for fixed sample size n it is irrelevant whether we view the critical values as being constant or as depending on n. Hence, the conclusions regarding the finite-sample distributions of PMSEs drawn in the previous paragraph carry over to the case of consistent model selection. When it comes to the *pointwise* asymptotic distribution of PMSEs based on *consistent* model selection procedures a difference arises: It is easy to see that the pointwise asymptotic distribution of $n^{1/2}(\tilde{\eta} - \eta)$ is then typically normal and coincides with the (pointwise) asymptotic distribution of $n^{1/2}(\hat{\eta}(M_0) - \eta)$, where $\hat{\eta}(M_0)$ is the infeasible 'oracle' that makes use of knowledge of the minimal true model M_0. This has been noted in Hannan and Quinn (1979) and Pötscher (1991, Lemma 1) who also issued a warning regarding the statistical interpretation of this result. Nevertheless, this property of PMSEs based on consistent model selection procedures has frequently – and incorrectly – been interpreted in the literature as saying that consistent model selection has no effect asymptotically on the distributional properties of the parameter estimator and that one can estimate θ asymptotically as efficient as if knowledge about the minimal true model were available: Two prominent examples are Bunea (2004) and Fan and Li (2001) who advertise this property of their estimators, the latter reference dubbing this property the 'oracle property'. However, this interpretation is a fallacy: The 'oracle property' is essentially a reincarnation of the 'superefficiency' phenomenon à la the 'superefficiency' of Hodges' estimator, and does not reflect actual statistical performance. [Other instances in the literature where this misleading interpretation has been reported include Geweke and Meese (1981), Lütkepohl (1990, p.120), Hidalgo (2002), Hall and Peixe (2003), Dufour et al. (2006).] Mathematically speaking, the problem is that convergence of the finite-sample distributions to their asymptotic counterparts is highly non-uniform in the parameter, and that the 'oracle property' results are only *pointwise* asymptotic results. This has already been noted by Shibata (1986a) and Kabaila (1995, 1996). While pointwise asymptotics are unable to capture the effects of consistent model selection, a more appropriate asymptotic analysis that allows the true parameter to depend on sample size very well reveals these effects; see Leeb and Pötscher (2005) for an extensive discussion. The non-uniformity in the convergence of the finite-sample distributions is of course related to the unboundedness of the maximal risk, cf. (19), discussed in the previous subsection.

Estimation of the distribution of PMSEs. As discussed above, the finite-sample (as well as the asymptotic) distribution of PMSEs typically depend on unknown parameters in a complicated fashion. In order to be able to utilize these distributions for inference one has to estimate these distributions. While consistent estimators for the distribution of PMSEs can be constructed, it has been shown in Leeb and Pötscher (2006b, 2008b) that such estimators are necessarily of low quality in the sense that no estimator can be uniformly consistent. Such 'impossibility' results also arise for a large class of shrinkage-type estimators, see Leeb and Pötscher (2006a), Pötscher and Leeb (2007),

and Pötscher and Schneider (2007). See also Section 2.3 in Leeb and Pötscher (2005) for a simple exposition of the issues involved here.

Confidence sets post model selection. Problems related to the constructions of valid confidence intervals post model selection are discussed in Kabaila (1995, 1998), Pötscher (1995), Kabaila and Leeb (2006), Leeb (2007), and Pötscher (2007).

3 Model Selection in Large- or Infinite-Dimensional Models

In Section 2 we have mainly concentrated on the case where there exists a true model in the set of candidate models $\mathcal{M}$ and where the cardinality of $\mathcal{M}$ is finite and independent of sample size n. It can, however, be argued that the need for model selection is particularly great when the dimension of the candidate models (e.g., number of potentially important explanatory variables) is large in relation to sample size and/or no model in $\mathcal{M}$ is correct (i.e., the models fitted to the data constitute only an approximation to the data generating process). To analyze a scenario like this, it is often more appropriate to assume that the true data-generating process is infinite-dimensional, and that one tries to identify a 'good' finite-dimensional model on the basis of the data.

Throughout this section, we consider the following 'infinite-dimensional' extension of Example 1: In the setting of that example, assume that the number of regressors, i.e., K in (1), is infinite. To ensure that the response Y and its mean $X\theta$ are well-defined, assume that θ as well as the rows-vectors of X are square-summable, i.e., θ and $X'_{i,\cdot}$ are in l_2. Moreover, assume that the infinite-dimensional 'matrices' (i.e., operators on l_2) $X'X/n$ converge in the operator norm to an invertible operator Q as $n \to \infty$. We also make the assumption that the vector of errors u is distributed as $N(0, \sigma^2 I_n)$, $0 < \sigma^2 < \infty$. Given a sample of size n, consider model selection from a family $\mathcal{M}_n$ of finite-dimensional candidate models. Throughout, we always assume that each model $M \in \mathcal{M}_n$ is such that the $n \times k_M$ matrix of those explanatory variables included in the model M has full column-rank k_M; we make this assumption for convenience, although it is not necessary for all the results discussed below. The collection of candidate models $\mathcal{M}_n$ considered at sample size n is assumed to be finite or countable, and it is allowed to depend on sample size satisfying $\mathcal{M}_n \subseteq \mathcal{M}_{n+1}$ (although this is again not necessary for all the results discussed below). This setting is sufficiently general to present the relevant results while it is sufficiently simple to keep the notation and assumptions manageable. We refer to the literature for more general results.

One of the early analyses of model selection in infinite-dimensional models is Shibata (1980); see also Shibata (1981a,b). In essence, these papers establish that the PMSE based on AIC (or FPE) is pointwise asymptotically loss-efficient *provided* that the true data-generating process is infinite-dimensional

(and that the collection of candidate models $\mathcal{M}_n$ increases appropriately with n). In the setting of (1) with $K = \infty$ as introduced above, define the loss $L_n(\theta, \bar{\theta})$ of an estimator $\bar{\theta}$ of θ (taking values in l_2) as

$$L_n(\theta, \bar{\theta}) = (\bar{\theta} - \theta)' \frac{X'X}{n} (\bar{\theta} - \theta),$$

and let $R_n(\theta, \bar{\theta}) = \mathbb{E}_{n,\theta}(L_n(\theta, \bar{\theta}))$ denote the corresponding risk. Given a model selection procedure $\hat{M}$, Shibata (1981b) compares the loss of the PMSE $\hat{\theta}(\hat{M})$, i.e., $L_n(\theta, \hat{\theta}(\hat{M}))$ with the minimum of the losses of the least-squares estimators $\hat{\theta}(M)$ corresponding to all the models M in $\mathcal{M}_n$, i.e., with $\inf_{M \in \mathcal{M}_n} L_n(\theta, \hat{\theta}(M))$. If $\hat{M}_{AIC}$ is chosen by the minimum AIC method, i.e., $\hat{M}_{AIC}$ is a (measurable) minimizer of $\mathsf{AIC}_n(M)$ over $M \in \mathcal{M}_n$, then Shibata (1981b) shows that

$$\frac{L_n(\theta, \hat{\theta}(\hat{M}_{AIC}))}{\inf_{M \in \mathcal{M}_n} L_n(\theta, \hat{\theta}(M))} \xrightarrow{p} 1 \tag{21}$$

provided that the true parameter θ is truly infinite-dimensional (i.e., has infinitely many non-zero coordinates), provided that the candidate models are nested in the sense that $\mathcal{M}_n = \{M(p) : 0 \leq p \leq K_n\}$, and provided that the number of candidate models $K_n + 1$ satisfies $K_n \to \infty$ and $K_n = o(n)$. [The model $M(p)$ here refers to the model containing the first p regressors; cf. Example 1.] Relation (21) continues to hold if in the denominator loss is replaced by risk. All these results carry over if, instead of the AIC criterion, either Mallows' C_p or FPE are used for model selection. [Cf. Theorems 2.1, 2.2, 3.1, and the discussion leading up to Assumption 1, as well as Section 5 in Shibata (1981b).] Shibata (1981b) points out that (21) does not hold for model selectors based on BIC-type model selection criteria. The results in Shibata (1981b) in fact allow for classes of candidate models more general than the nested case considered here, provided that the condition that θ is infinite-dimensional is replaced by a more complicated condition that, in essence, requires that the candidate models considered at sample size n do not fit the true data-generating process 'too well'; see Assumptions 1 and 2 in Shibata (1981b). For order selection in autoregressive models approximating an infinite-order autoregressive data generating process, results similar to those just presented are given by Shibata (1980); here models are evaluated in terms of their predictive performance out-of-sample, where the model selection and fitting step on the one hand and the prediction step on the other hand are based on two independent realizations of the same time series. Recently, Ing and Wei (2005) have obtained parallel results for the case where one and the same realization of the process is used for the estimation, selection, and prediction step. Shibata (1981a) also considers selection of approximating autoregressive models where the models are now evaluated by the performance of the corresponding estimate of the spectral density (at a fixed frequency). Pointwise asymptotic loss efficiency results in the above sense are also established in Li (1987), Polyak and Tsybakov (1990), and Shao (1997) for a variety of

other methods including generalized cross-validation and cross-validation, and under somewhat different sets of assumptions. See also Breiman and Friedman (1983) for a similar result about the criterion (12) when models are evaluated by their predictive performance out-of-sample.

All the pointwise asymptotic loss-efficiency results for conservative model selection procedures like (21) mentioned above rely on the central assumption that the true data-generating process is 'not too well approximated' by the finite-dimensional candidate models considered at sample size n as $n \to \infty$. In the simple setting considered in (21), this is guaranteed by assuming that θ is infinite-dimensional, and, in more general settings, by conditions like Assumption 2 in Shibata (1981b). If that central assumption is violated, statements like (21) will typically break down for conservative model selection procedures (like AIC or FPE). In fact, Shao (1997) considers a scenario where BIC and related consistent model selection procedures are pointwise asymptotically loss-efficient when the true model is finite-dimensional. [This is in line with the discussion of the 'oracle' phenomenon in Section 2.2 for the case of finitely many candidate models.] These findings suggest a dichotomy: If the true model is infinite-dimensional, conservative procedures like AIC are pointwise asymptotically loss-efficient while consistent procedures like BIC are not; if the true model is finite-dimensional, the situation is reversed (under appropriate assumptions). However, one should not read too much into these results for the following reasons: The true model may be infinite-dimensional, suggesting an advantage of AIC or a related procedure over BIC. At a given sample size, however, one of the finite-dimensional candidate models may provide a very good approximation to the true data-generating process, and hence the pointwise asymptotic loss-efficiency result favoring AIC may not be relevant. Conversely, the true model may be finite-dimensional, suggesting an advantage of consistent procedures like BIC, but, compared to the given sample size, some of the non-zero parameters may be moderately small, thereby fooling the consistent model selection procedure into choosing an incorrect model which then translates into bad risk behavior of the PMSE. Mathematically speaking, the problem is that the asymptotic loss-efficiency results discussed above are only pointwise results and do not hold uniformly w.r.t. the parameter θ: Kabaila (2002), shows, in a simple setting where (21) holds, that for given sample size n, there exists a parameter $\theta = \theta_n$ with infinitely many non-zero coordinates such that

$$\frac{L_n(\theta, \hat{\theta}(\hat{M}_{AIC}))}{L_n(\theta, \hat{\theta}(\hat{M}_{BIC}))} \geq 1,$$

and such that the ratio on the left-hand side in the above display is greater than 2 with probability larger than 0.13. This shows that the results of Shibata (1981b), which entail that

$$\limsup_{n\to\infty} L_n(\theta, \hat{\theta}(\hat{M}_{AIC}))/L_n(\theta, \hat{\theta}(\hat{M}_{BIC})) \leq 1$$

for every *fixed* θ as $n \to \infty$, do not hold uniformly in θ (as for fixed n there exists a parameter θ for which the situation is reversed). Similarly, Shao's (1997) asymptotic loss-efficiency results for consistent procedures mentioned above are only pointwise asymptotic results and thus similarly problematic. [Compare with the discussion of the 'oracle' phenomenon in Section 2.2 showing that consistent model selection procedures have bad maximal risk properties when selecting from a finite family of finite-dimensional models.]

Given the dichotomy arising from the *pointwise* asymptotic loss-efficiency results, attempts have been made to devise 'adaptive' model selection procedures that work well in both scenarios, i.e., procedures that combine the beneficial properties of both consistent and conservative procedures but avoid their detrimental properties. While this can be achieved in a *pointwise* asymptotic framework (Yang (2007), Ing (2007)), it is not surprising – given the preceding discussion – that it is impossible to achieve this goal in a uniform sense: No model selection procedure can simultaneously be consistent (like BIC) and minimax-rate adaptive (like AIC) as shown in Yang (2005). [This is related to the fact that consistent model selection procedures lead to PMSEs that have maximal risk that diverges to infinity as sample size increases, even for finite dimensional models; see the discussion regarding (19) in Section 2.2.]

The discussion so far again demonstrates that *pointwise* large-sample limit analyses of model selection procedures can paint a picture that is misleading in the sense that it need not have much resemblance to the situation in finite samples of *any* size. We now turn to two recent lines of research that analyze model selection procedures from a different perspective. Both of these lines of research rely on a combination of finite-sample results and asymptotic results that hold uniformly over (certain regions of) the parameter space, instead of pointwise asymptotic analyses.

In recent years, finite-sample risk bounds for PMSEs have been developed in considerable generality; see Barron and Cover (1991), Barron, Birgé and Massart (1999), and the references given in that paper. The following results are adapted from Birgé and Massart (2001) to our setting. Assume that the error variance σ^2 is known. For a (finite or countable) collection $\mathcal{M}_n$ of candidate models, consider the model selector $\hat{M}$ that minimizes the following C_p-like criterion over $M \in \mathcal{M}_n$:

$$\mathsf{crit}(M) = RSS(M) \; + \; \kappa(1 + \sqrt{2l_M})^2 \sigma^2 k_M.$$

This criterion depends on the user-specified constants $\kappa > 1$ and $l_M \geq 0$ that are chosen so that

$$\sum_{\substack{M \in \mathcal{M}_n, \\ k_M > 0}} e^{-l_M k_M} \leq \Sigma < \infty. \tag{22}$$

[For $\kappa(1 + \sqrt{2l_M})^2 = 2$ this criterion coincides with (6) up to an irrelevant additive constant.] Then the resulting post-model-selection estimator $\hat{\theta}(\hat{M})$ has a risk $R_n(\theta, \hat{\theta}(\hat{M})) = \mathbb{E}_{n,\theta}[L_n(\theta, \hat{\theta}(\hat{M}))]$ satisfying

$$R_n(\theta, \hat{\theta}(\hat{M})) \leq C(\kappa) n^{-1} \left[\inf_{M \in \mathcal{M}_n} \left(||(I - P_M)X\theta||^2 + \sigma^2 k_M (1 + l_M) \right) + \sigma^2 \Sigma \right], \tag{23}$$

for $\theta \in l_2$, where the constant $C(\kappa)$ is given by $C(\kappa) = 12\kappa(\kappa+1)^3/(\kappa-1)^3$. [Cf. Theorem 2 and (3.12) in Birgé and Massart (2001), and observe that $(1 + \sqrt{2x})^2 \leq 3(1 + x)$ for $x \geq 0$.] We stress that the upper bound in (23) holds under no additional assumptions on the unknown parameter θ other than $\theta \in l_2$. The upper bound in (23) equals $C(\kappa)$ times the sum of two terms: The first one is the infimum over all candidate models of a 'penalized' version of the bias of the model M, i.e., $||(I-P_M)X\theta||^2/n$, where the 'penalty' is given by $\sigma^2 k_M(1 + l_M)/n$. It should be noted that $R_n(\theta, \hat{\theta}(M))$, i.e., the risk when fitting model M, is given by $R_n(\theta, \hat{\theta}(M)) = ||(I - P_M)X\theta||^2/n + \sigma^2 k_M/n$. If, in addition, the constants l_M can be chosen to be bounded, i.e., $l_M \leq L$ for each $M \in \mathcal{M}_n$, while still satisfying (22), it follows from (23) that

$$R_n(\theta, \hat{\theta}(\hat{M})) \leq (1 + L)C(\kappa) \left[\inf_{M \in \mathcal{M}_n} R_n(\theta, \hat{\theta}(M)) + \frac{\sigma^2}{n} \Sigma \right]. \tag{24}$$

Provided that $l_M \leq L$ for each $M \in \mathcal{M}_n$, we hence see that the risk of the post-model-selection estimator $\hat{\theta}(\hat{M})$ is bounded by a constant multiple of the risk of the minimal-risk candidate model plus a constant. Suppose that one of the finite-dimensional candidate models, say, M_0, is a correct model for θ, and that M_0 is the smallest model with that property. Then $\inf_{M \in \mathcal{M}_n} R_n(\theta, \hat{\theta}(M))$ is not larger than $\sigma^2 k_{M_0}/n$; in that case, the infimum in (24) is of the same order as $\sigma^2\Sigma/n$. Conversely, suppose that θ is infinite-dimensional. Then $\inf_{M \in \mathcal{M}_n} R_n(\theta, \hat{\theta}(M))$ typically goes to zero slower than $1/n$, to the effect that $\inf_{M \in \mathcal{M}_n} R_n(\theta, \hat{\theta}(M))$ is now the dominating factor in the upper bound in (24). Birgé and Massart (2001) argue that, without additional assumptions on the true parameter θ, the finite-sample upper bound in (24) is qualitatively best possible (cf. the discussion leading up to (2.9) in that paper); see also Sections 3.3.1 and 3.3.2 of that paper for a discussion of the choice of the constants κ and l_M in relation to the family of candidate models $\mathcal{M}_n$. It can furthermore be shown that the maximal risk of $\hat{\theta}(\hat{M})$ over certain regions Θ in the parameter space is not larger than a constant times the minimax risk over Θ, i.e.,

$$\sup_{\theta \in \Theta} R_n(\theta, \hat{\theta}(\hat{M})) \leq C(\Theta, \kappa, L) \inf_{\bar{\theta}} \sup_{\theta \in \Theta} R_n(\theta, \bar{\theta}), \tag{25}$$

where the infimum is taken over all estimators $\bar{\theta}$, and where the constant $C(\Theta, \kappa, L)$ depends on the indicated quantities but not on sample size. Results of that kind hold, for example, in case the candidate models are the nested models $M(p)$ and the parameter set $\Theta \subseteq l_2$ is, after an appropriate re-parameterization, a Sobolev or a Besov body (cf. Section 6 of Birgé and Massart (2001) or Section 5 of Barron, Birgé and Massart (1999)). We also note that the results of Barron, Birgé and Massart (1999) are more general

and cover the Gaussian regression model discussed here as a special case; similar results continue to hold for other problems including maximum likelihood density estimation, minimum $\mathbb{L}_1$ regression and general projection estimators. For further results in that direction, see Barron (1991), Barron (1998), Yang and Barron (1998), and Yang (1999). Furthermore, Birgé (2006) derives results similar to (23)–(25) for model selection based on preliminary tests. He finds that the resulting PMSEs sometimes perform favorably compared to the estimators based on penalized maximum likelihood or penalized least-squares considered above, but that the implementation of the preliminary tests can be difficult.

Risk bounds like (23)-(25) above are sometimes called 'oracle inequalities' in the literature (although there is no precise definition of this term). Informally, these bounds state that the risk of the PMSE is 'not too far away' from the 'risk-target', i.e., from the risk corresponding to the model or to the estimator that an all-seeing oracle would choose.

Beran (1996) also studies the loss of PMSEs, but has a different focus. Instead of concentrating on oracle inequalities for the risk, that paper studies the problem of estimating the risk or loss of PMSEs; see also Kneip (1994), Beran and Dümbgen (1998), and Beran (2000). For the sake of simplicity, consider as the collection of candidate models at sample size n the set of all nested models of order up to n, i.e., $\mathcal{M}_n = \{M(p) : 0 \leq p \leq n\}$, assume again that the variance σ^2 is known, and let $\hat{M}_{C_p}$ denote a (measurable) minimizer of the Mallows' C_p objective function

$$\mathsf{MC}_n(M) = RSS(M) + 2k_M\sigma^2 - n\sigma^2$$

over the set of candidate models. It then follows from Theorem 2.1 and Example 3 of Beran and Dümbgen (1998) that

$$\mathbb{E}_{n,\theta}\left|L_n(\theta, \hat{\theta}(\hat{M}_{C_p})) - \inf_{M\in\mathcal{M}_n} L_n(\theta, \hat{\theta}(M))\right| \leq \frac{C}{\sqrt{n}}\left(\sigma^2 + \sigma\sqrt{\theta'\frac{X'X}{n}\theta}\right) \quad (26)$$

and

$$\mathbb{E}_{n,\theta}\left|L_n(\theta, \hat{\theta}(\hat{M}_{C_p})) - n^{-1}\mathsf{MC}_n(\hat{M}_{C_p})\right| \leq \frac{C}{\sqrt{n}}\left(\sigma^2 + \sigma\sqrt{\theta'\frac{X'X}{n}\theta}\right), \quad (27)$$

where C is a constant independent of n, X, θ, and σ^2. The relation (26) is similar to (24) in that it relates the selected model to the best model. [Of course, the results in (24) and (26) are qualitatively different, but we shall not discuss the differences here. Beran and Dümbgen (1998) also provide a bound similar to (26) for the risk instead of the loss; moreover, they show that an extension of $\hat{\theta}(\hat{M}_{C_p})$, which is based on smooth shrinkage, is asymptotically minimax over certain regions in parameter space.] The result in (27) differs from those discussed so far in the sense that it shows that the loss of the model selected by

Mallows' C_p, i.e., $L_n(\theta, \hat{\theta}(\hat{M}_{C_p}))$, which is unknown in practice, can actually be estimated by n^{-1} times the value of the C_p objective function $\mathsf{MC}_n(\hat{M}_{C_p})$, provided only that $\sqrt{n}$ is large in relation to $\sigma^2 + \sigma(\theta' X' X \theta/n)^{1/2}$. [Note that the upper bounds in (26) and (27), although unknown, can be estimated, because $\theta' X' X \theta/n$ can be estimated. The error variance σ^2 is assumed to be known here; in practice, σ^2 can often be estimated with reasonable accuracy.] A similar statement also holds for the risk corresponding to the model selected by C_p. The ability to actually estimate the risk or loss of the PMSE is important, because it allows for inference after model selection, like, e.g., the construction of asymptotically valid confidence balls. See Beran (1996), Beran and Dümbgen (1998), and Beran (2000) for results in that direction. We also note that these papers allow for more general classes of candidate models and also for more general estimators, that is, for estimators based on smooth shrinkage.

So far, we have considered the fixed-design setting, i.e., the regressors were assumed to be non-random. The case of random design is studied by Baraud (2002), Wegkamp (2003), and Birgé (2004), based on the results of Barron, Birgé and Massart (1999). Leeb (2006b) gives results similar to (26) and (27) in the case of random design, where the loss is defined as squared-error loss for out-of-sample prediction, when a variant of the generalized cross-validation criterion (or the criterion (12)) is used for model selection. Predictive inference after model selection is studied in Leeb (2007).

4 Related Procedures Based on Shrinkage and Model Averaging

Classical shrinkage-type estimators include the James-Stein estimator and related methods (cf. James and Stein (1961), Strawderman and Cohen (1971)), or the ridge estimator (cf. Hoerl and Kennard (1970)). In recent years, there has been a renewed interest in shrinkage-type estimators; examples include the bridge estimator of Frank and Friedman (1993), the nonnegative garrote of Breiman (1995), the LASSO of Tibshirani (1996), the lasso-type estimators analyzed by Knight and Fu (2000), the smoothly clipped absolute deviation (SCAD) estimators proposed by Fan and Li (2001), or the adaptive LASSO of Zou (2006). Many of these estimators just mentioned are instances of penalized maximum likelihood or least squares estimators.

Model averaging estimators – instead of selecting one candidate model and the corresponding estimator – form a weighted sum of the estimators corresponding to each of the candidate models where the weights typically are allowed to depend on the data. Model averaging estimators occur naturally in a Bayesian framework, where each model is weighted by its posterior probability. Good entry points into the considerable amount of literature on Bayesian model averaging are Hoeting, Madigan, Raftery, and Volinsky

(1999), or Brown, Vannucci and Fearn (2002); see also the references given in these papers. Of course, model averaging methods have also been analyzed from a frequentist perspective; see, for example, Buckland, Burnham, and Augustin (1997), Magnus (2002), Juditsky and Nemirovski (2000), Yang (2001, 2003), Hjort and Claeskens (2003), Danilov and Magnus (2004), Leung and Barron (2006), as well as Bunea, Tsybakov and Wegkamp (2007).

Both shrinkage-type estimators and model averaging estimators can be regarded as extensions of PMSEs: Clearly, PMSEs can be viewed as special cases of shrinkage-type estimators, in the sense that PMSEs restrict (shrink) certain individual components of the parameter vector to zero. PMSEs can also be viewed as a particular case of model averaging estimators, where the weights are such that the selected model gets weight one and the other models get weight zero; cf. (2). This suggests that a number of phenomena that one can observe for PMSEs have counterparts in the larger class of shrinkage estimators or the class of estimators based on model averaging: Certain shrinkage estimators like the SCAD of Fan and Li (2001) or the adaptive LASSO of Zou (2006) can have a 'sparsity property' in the sense that zero components of the true parameter are estimated as exactly zero with probability approaching one as sample size increases (provided that the estimator's tuning parameter is chosen appropriately); consistent PMSEs have the same property. It is therefore not surprising that shrinkage estimators that have this 'sparsity property' perform unfavorably in terms of worst-case risk in large samples (cf. the discussion in Section 2.2). In particular, the worst-case risk of shrinkage estimators that have the sparsity property increases to infinity with sample size; cf. Leeb and Pötscher (2008a). Also, the phenomena discussed in Section 2.3, namely that the distribution is typically highly non-normal and that the cdf of PMSEs can not be estimated with reasonable accuracy, also occur with shrinkage estimators or model averaging estimators; cf. Leeb and Pötscher (2006a), Pötscher (2006), Pötscher and Leeb (2007), and Pötscher and Schneider (2007).

5 Further Reading

Apart from the expository articles already mentioned (Hocking (1976), Thompson (1978a,b), Amemiya (1980), Giles and Giles (1993), Hansen and Yu (2001), Rao and Wu (2001), Leeb and Pötscher (2005)), the article DeGooijer et al. (1985) provides a survey of model selection in time series analysis; see also Chapter 5 of Hannan and Deistler (1988). The Bayesian approach to model selection is discussed in Hoeting et al. (1999) and Berger and Pericchi (2001).

The books by Judge and Bock (1978), Linhart and Zucchini (1986), Choi (1992), McQuarrie and Tsai (1998), Burnham and Anderson (2002), Miller (2002), Saleh (2006), and Konishi and Kitagawa (2008) deal with various aspects of model selection.

There is also a considerable body of literature on model selection and related methods in the areas of machine learning and empirical risk minimization, mainly focusing on classification and pattern recognition problems; see, e.g., Boucheron, Bousquet, and Lugosi (2005), and Cesa-Bianchi and Lugosi (2006).

We finally mention a development that circles around the idea of automated discovery and automated modeling; for an introduction see Phillips (2005) and references therein.

6 References

Ahmed, S. E. & A. K. Basu (2000): Least squares, preliminary test and Stein-type estimation in general vector AR(p) models. *Statistica Neerlandica* 54, 47–66.

Akaike, H. (1969): Fitting autoregressive models for prediction. *Annals of the Institute for Statistical Mathematics* 21, 243-247.

Akaike, H. (1970): Statistical predictor identification. *Annals of the Institute for Statistical Mathematics* 22, 203-217.

Akaike, H. (1973): Information theory and an extension of the maximum likelihood principle. In: B.N. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory.* Akadémiai Kiadó, Budapest.

Allen, D. M. (1971): Mean square error of prediction as a criterion for selecting variables. *Technometrics* 13, 469-475.

Allen, D. M. (1974): The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16, 125-127.

Amemiya, T. (1980): Selection of regressors. *International Economic Review* 21, 331-354.

An, H. Z. & Z. G. Chen (1986): The identification of ARMA processes. *Journal of Applied Probability* Special Vol. 23A, 75–87.

An, H. Z. & L. Gu (1985): On the selection of regression variables. *Acta Mathematicae Applicatae Sinica* 2, 27-36.

Anderson, T. W. (1962): The choice of the degree of a polynomial regression as a multiple decision problem. *Annals of Mathematical Statistics* 33, 255–265.

Anderson, T. W. (1963): Determination of the order of dependence in normally distributed time series. In: M. Rosenblatt (ed.), *Time Series Analysis*, 425–446. Wiley, New York.

Bancroft, T. A. & C. P. Han (1977): Inference based on conditional specification: A note and a bibliography. *International Statistical Review* 45, 117–127.

Baraud, Y. (2002): Model selection for regression on a random design. *ESAIM Probability and Statistics* 6, 127–146.

Barron, A. R. (1991): Complexity regularization with application to artificial neural networks. In: Nonparametric functional estimation and related

topics (Spetses, 1990), *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* 335, Kluwer, Dordrecht, 561–576.

Barron, A. R. (1999): Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In: *Bayesian Statistics*, 6 (Alcoceber, 1998), Oxford University Press, New York, 27-52.

Barron, A. R., Birgé, L. & P. Massart (1999): Risk bounds for model selection via penalization. *Probability Theory and Related Fields* 113, 301–413.

Barron, A. R. & T. M. Cover (1991): Minimum complexity density estimation. *IEEE Transactions on Information Theory* 37, 1034-1054. (Corrections: *IEEE Transactions on Information Theory* 37, 1738.)

Bauer, P., Pötscher, B. M. & P. Hackl (1988): Model selection by multiple test procedures. *Statistics* 19, 39–44.

Beran, R. (1996): Confidence sets centered at C_p-estimators. *Annals of the Institute of Statistical Mathematics* 48, 1–15.

Beran, R. (2000): REACT scatterplot smoothers: Superefficiency through basis economy. *Journal of the American Statistical Association* 95, 155–171.

Beran, R. & L. Dümbgen (1998): Modulation of estimators and confidence sets. *Annals of Statistics* 26, 1826–1856.

Berger, J. O. & L. R. Pericchi (2001): Objective Bayesian methods for model selection: Introduction and comparison. In: P. Lahiri (ed.), Model Selection. *IMS Lecture Notes Monograph Series* Vol. 38, 135–193.

Bhansali, R. J. (1999). Parameter estimation and model selection for multistep prediction of a time series: A review. In: Subir Ghosh (ed.), *Asymptotics, Nonparametrics and Time Series – A Tribute to Madan Lal Puri*, Marcel Dekker, New York, 201-225.

Bhansali, R. J. & D. Y. Downham (1977): Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika* 64, 547-551.

Birgé, L. (2004): Model selection for Gaussian regression with random design. *Bernoulli* 10, 1039–1051.

Birgé, L. (2006): Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Annales Institute Henri Poincaré – PR* 42, 273–325.

Birgé, L. & P. Massart (2001): Gaussian model selection. *Journal of the European Mathematical Society* 3, 203–268.

Breiman, L. (1995): Better subset regression using the nonnegative garrote. *Technometrics* 37, 373–384.

Breiman, L. & D. Freedman (1983): How many variables should be entered in a regression equation? *Journal of the American Statistical Association* 78, 131–136.

Brook, R. J. (1976): On the use of a regret function to set significance points in prior tests of estimation. *Journal of the American Statistical Asso-*

ciation 71, 126-131. (Correction: *Journal of the American Statistical Association* 71, 1010.)

Brown, P. J., Vannucci, M. & T. Fearn (2002): Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society B* 64, 519-536.

Boucheron, S., Bousquet, O. & G. Lugosi (2005): Theory of classification: A survey of some recent advances. *ESAIM Probability and Statistics* 9, 323–375.

Buckland S. T., Burnham, K. P. & N. H. Augustin (1997): Model selection: An integral part of inference. *Biometrics* 53, 603–618.

Bunea, F. (2004): Consistent covariate selection and post model selection inference in semiparametric regression. *Annals of Statistics* 32, 898-927.

Bunea, F., Tsybakov, A. & M. H. Wegkamp (2007): Aggregation for Gaussian regression. *Annals of Statistics* 35, 1674-1697.

Bunea, F., M. H. Wegkamp & A. Auguste (2006): Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference* 136, 4349-4364.

Burnham, K. P. & D. R. Anderson (2002): *Model Selection and Multimodal Inference (2nd edition).* New York: Springer.

Cesa-Bianchi, N. & G. Lugosi (2006): *Prediction, learning, and games.* Cambridge University Press, Cambridge.

Chen, S. S., Donoho, D. L.& M. A. Saunders (1998): Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* 20, 33-61.

Chen, Z. G. & J. Y. Ni (1989): Subset regression time series and its modeling procedures. *Journal of Multivariate Analysis* 31, 266-288.

Choi, B. (1992): *ARMA Model Identification.* New York: Springer.

Claeskens, G. & N. L. Hjort (2003): The focused information criterion. *Journal of the American Statistical Association* 98, 900-916.

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

Danilov, D. & J. R. Magnus (2004): On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122, 27-46.

Davisson, L. D. (1965): The prediction error of stationary Gaussian time series of unknown covariance. *IEEE Transactions on Information Theory* 11, 527-532.

DeGooijer, J. G., Bovas, A., Gould, A. & L. Robinson (1985): Methods for determining the order of an autoregressive-moving average process: A survey. *International Statistical Review* 53, 301-329.

Dijkstra, T. K. & J. H. Veldkamp (1988): Data-driven selection of regressors and the bootstrap. In: T. K. Dijkstra (ed.), *Lecture Notes in Economics and Mathematical Systems* 307, 17–38.

Draper, N. R. & H. Smith (1981): *Applied Regression Analysis* (2nd edition). New York: Wiley.

Droge, B. (1993): On finite-sample properties of adaptive least squares regression estimates. *Statistics* 24, 181-203.

Droge, B. & T. Georg (1995): On selecting the smoothing parameter of least squares regression estimates using the minimax regret approach. *Statistics and Decisions* 13, 1-20.

Dufour, J. M., Pelletier, D. & E. Renault (2006): Short run and long run causality in time series: Inference. *Journal of Econometrics* 132, 337-362.

Fan, J. & R. Li (2001): Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348-1360.

Findley, D. F. (1985): On the unbiasedness property of AIC for exact or approximating linear stochastic time series models. *Journal of Time Series Analysis* 6, 229-252.

Findley, D. F. (1991): Model selection for multistep-ahead forecasting. *Amer. Stat. Assoc. Proc. Bus. Econ. Stat. Sec.* 243-247.

Findley, D. F. & C. Z. Wei (2002): AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autoregressions and related models. *Journal of Multivariate Analysis* 83, 415-450.

Foster, D. P. & E. I. George (1994): The risk inflation criterion for multiple regression. *Annals of Statistics* 22, 1947-1975.

Frank, I. E. & J. H. Friedman (1993): A statistical view of some chemometrics regression tools (with discussion). *Technometrics* 35, 109-148.

Francq, C., Roussignol, M. & J. M. Zakoïan (2001): Conditional heteroskedasticity driven by hidden Markov chains. *Journal of Time Series Analysis* 22, 197-220.

George, E. I. & D. P. Foster (2000): Calibration and empirical Bayes variable selection. *Biometrika* 87, 731-747.

Geweke, J. & R. Meese (1981): Estimating regression models of finite but unknown order. *International Economic Review* 22, 55-70.

Giles, J. A. & D. E. A. Giles (1993): Pre-test estimation and testing in econometrics: recent developments. *Journal of Economic Surveys* 7, 145-197.

Guyon, X. & J. Yao (1999): On the underfitting and overfitting sets of models chosen by order selection criteria. *Journal of Multivariate Analysis* 70, 221–249.

Hall, A. R. & F. P. M. Peixe (2003): A consistent method for the selection of relevant instruments. *Econometric Reviews* 22, 269-287.

Hannan, E. J. (1980): The estimation of the order of an ARMA process. *Annals of Statistics* 8, 1071-1081.

Hannan, E. J. (1981): Estimating the dimension of a linear system. *Journal of Multivariate Analysis* 11, 459–473.

Hannan, E. J. & M. Deistler (1988): *The Statistical Theory of Linear Systems*. New York: Wiley.

Hannan, E. J. & B. G. Quinn (1979): The determination of the order of an autoregression. *Journal of the Royal Statistical Society B* 41, 190-195.

Hansen, M. H. & B. Yu (2001): Model selection and the principle of minimum description length. *Journal of the American Statistical Association* 96, 746–774.

Haughton, D. (1991) Consistency of a class of information criteria for model selection in nonlinear regression. *Communications in Statistics. Theory and Methods* 20, 1619–1629.

Hemerly, E. M. & M.H.A. Davis (1989): Strong consistency of the PLS criterion for order determination of autoregressive processes. *Annals of Statistics* 17, 941-946.

Hidalgo, J. (2002): Consistent order selection with strongly dependent data and its application to efficient estimation. *Journal of Econometrics* 110, 213-239.

Hjort, N. L. & G. Claeskens (2003): Frequentist model average estimators. *Journal of the American Statistical Association* 98, 879-899.

Hocking, R. R. (1976): The analysis and selection of variables in linear regression. *Biometrics* 32, 1-49.

Hoerl, A. E. & R. W. Kennard (1970): Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55-67.

Hoeting, J. A., Madigan, D., Raftery, A. & C. T. Volinsky (1999): Bayesian model averaging: A tutorial. *Statistical Science* 14, 382-401. (Corrections: *Statistical Science* 15, 193-195.)

Hosoya, Y. (1984): Information criteria and tests for time series models. In: O. D. Anderson (ed.), *Time Series Analysis: Theory and Practice* 5, 39-52. Amsterdam: North-Holland.

Hosoya, Y. (1986): A simultaneous test in the presence of nested alternative hypotheses. *Journal of Applied Probability* Special Vol. 23A, 187–200.

Hurvich, M. M. & C. L. Tsai (1989): Regression and time series model selection in small samples. *Biometrika* 76, 297-307.

Ing, C. K. (2004): Selecting optimal multistep predictors for autoregressive processes of unknown order. *Annals of Statistics* 32, 693-722.

Ing, C. K. (2007): Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Annals of Statistics* 35, 1238-1277.

Ing, C. K. & C. Z. Wei (2005): Order selection for same-realization predictions in autoregressive processes. *Annals of Statistics* 33, 2423–2474.

Ing, C. K. & C. Z. Wei (2006): A maximal moment inequality for long range dependent time series with applications to estimation and model selection. *Statistica Sinica* 16, 721–740.

James, W. and Stein, C. (1961): Estimation with quadratic loss. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, California University Press, Berkeley CA, 361–379.

Judge, G. G. & M. E. Bock (1978): *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*. Amsterdam: North-Holland.

Juditsky, A. & A. Nemirovski (2000): Functional aggregation for nonparametric regression. *Annals of Statistics* 28, 681–712.

Kabaila, P. (1995): The effect of model selection on confidence regions and prediction regions. *Econometric Theory* 11, 537–549.

Kabaila, P. (1996): The evaluation of model selection criteria: Pointwise limits in the parameter space. In: D. L. Dowe, K. B. Korb, and J. J. Oliver.(eds.), *Information, Statistics and Induction in Science*, 114-118. Singapore: World Scientific.

Kabaila, P. (1998): Valid confidence intervals in regression after variable selection. *Econometric Theory* 14, 463–482.

Kabaila, P. (2002): On variable selection in linear regression. *Econometric Theory* 18, 913–925.

Kabaila, P. & H. Leeb (2006): On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association* 101, 619-629.

Kempthorne, P. J. (1984): Admissible variable-selection procedures when fitting regression models by least squares for prediction. *Biometrika* 71, 593-597.

Kennedy, W. J. & T. A. Bancroft (1971): Model building for prediction in regression based upon repeated significance tests. *Annals of Mathematical Statistics* 42, 1273-1284.

Keribin, C. & D. Haughton (2003): Asymptotic probabilities of overestimating and under-estimating the order of a model in general regular families. *Communications in Statistics. Theory and Methods* 32, 1373–1404.

Kneip, A. (1994): Ordered linear smoothers. *Annals of Statistics* 22, 835-866.

Knight, K. (1989): Consistency of Akaike's information criterion for infinite variance autoregressive processes. *Annals of Statistics* 17, 824-840.

Knight, K. & W. Fu (2000): Asymptotics for lasso-type estimators. *Annals of Statistics* 28, 1356-1378.

Kohn, R. (1983): Consistent estimation of minimal subset dimension. *Econometrica* 51, 367-376.

Konishi, S. & G. Kitagawa (1996): Generalized information criteria in model selection. *Biometrika* 83, 875-890.

Konishi, S. & G. Kitagawa (2008): *Information Criteria and Statistical Modeling*. New York: Springer.

Leeb, H. (2005): The distribution of a linear predictor after model selection: Conditional finite-sample distributions and asymptotic approximations. *Journal of Statistical Planning and Inference* 134, 64–89.

Leeb, H. (2006a): The distribution of a linear predictor after model selection: Unconditional finite-sample distributions and asymptotic approximations. In: J. Rojo (ed.), *IMS Lecture Notes-Monograph Series* Vol. 49, 291-311.

Leeb, H. (2006b): Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. Manuscript, Department of Statistics, Yale University.

Leeb, H. (2007): Conditional predictive inference post model selection. Manuscript, Department of Statistics, Yale University.

Leeb, H. & B. M. Pötscher (2003): The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* 19, 100–142.

Leeb, H. & B. M. Pötscher (2005): Model selection and inference: Facts and fiction. *Econometric Theory* 21, 21–59.

Leeb, H. & B. M. Pötscher (2006a): Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory* 22, 69–97. (Corrigendum. *Econometric Theory*, forthcoming.)

Leeb, H. & B. M. Pötscher (2006b): Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* 34, 2554-2591.

Leeb, H. & B. M. Pötscher (2008a): Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* 142, 201–211.

Leeb, H. & B. M. Pötscher (2008b): Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* 24, forthcoming.

Leung, G. & A. R. Barron (2006): Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* 52, 3396–3410.

Li, K. C. (1987): Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* 15, 958-975.

Linhart, H. & W. Zucchini (1986): *Model Selection.* New York: Springer.

Lütkepohl, H. (1990): Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *Review of Economics and Statistics* 72, 116-125.

Magnus, J. R. (1999): The traditional pretest estimator. *Teoriya Veroyatnost. i Primenen.* 44, 401-418; translation in *Theory of Probability and Its Applications* 44 (2000), 293-308.

Magnus, J. R. (2002): Estimation of the mean of a univariate normal distribution with known variance. *The Econometrics Journal* 5, 225–236.

Mallows, C. L. (1965): Some approaches to regression problems. Unpublished manuscript.

Mallows, C. L. (1967): Choosing a subset regression. *Bell Telephone Laboratories*, unpublished report.

Mallows, C. L. (1973): Some comments on C_p. *Technometrics* 15, 661-675.

Mallows, C. L. (1995): More comments on C_p. *Technometrics* 37, 362-372.

McKay, R. J. (1977): Variable selection in multivariate regression: An application of simultaneous test procedures. *Journal of the Royal Statistical Society, Series B* 39, 371-380.

McQuarrie, A. D. R. & C. L. Tsai (1998): *Regression and time series model selection.* River Edge: World Scientific Publishing.

Miller, A. (2002): *Subset Selection in Regression (2nd edition).* Boca Raton: Chapman and Hall.

Nishii, R. (1984): Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* 12, 758-765.

Nishii, R. (1988): Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis* 27, 392–403.

Paulsen, J. (1984): Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis* 5, 115-127.

Phillips, P. C. B. (2005): Automated discovery in econometrics. *Econometric Theory* 21, 3–20.

Polyak, B. T. & A. B. Tsybakov (1990): Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory of Probability and Its Applications* 35, 293-306.

Pötscher, B. M. (1983): Order estimation in ARMA-models by Lagrangian multiplier tests. *Annals of Statistics* 11, 872-885.

Pötscher, B. M. (1985): The behaviour of the Lagrangian multiplier test in testing the orders of an ARMA-model. *Metrika* 32, 129-150.

Pötscher, B. M. (1989): Model selection under nonstationarity: Autoregressive models and stochastic linear regression models. *Annals of Statistics* 17, 1257-1274.

Pötscher, B. M. (1990): Estimation of autoregressive moving average order given an infinite number of models and approximation of spectral densities. *Journal of Time Series Analysis* 11, 165-179.

Pötscher, B. M. (1991): Effects of model selection on inference. *Econometric Theory* 7, 163–185.

Pötscher, B. M. (1995): Comment on 'The effect of model selection on confidence regions and prediction regions'. *Econometric Theory* 11, 550–559.

Pötscher, B. M. (2006): The distribution of model averaging estimators and an impossibility result regarding its estimation. In: H.-C. Ho, C.-K. Ing and T.-L. Lai (eds.), Time Series and Related Topics: In Memory of Ching-Zong Wei. *IMS Lecture Notes and Monograph Series* Vol. 52, 113-129.

Pötscher, B. M. (2007): Confidence sets based on sparse estimators are necessarily large. Working paper, Department of Statistics, University of Vienna, arXiv:0711.1036.

Pötscher, B. M. & H. Leeb (2007): On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. Working paper, Department of Statistics, University of Vienna, arXiv:0711.0660.

Pötscher, B. M. & A. J. Novak (1998): The distribution of estimators after model selection: Large and small sample results. *Journal of Statistical Computation and Simulation* 60, 19–56.

Pötscher, B. M. & U. Schneider (2007): On the distribution of the adaptive LASSO estimator. Working paper, Department of Statistics, University of Vienna.

Pötscher, B. M. & S. Srinivasan (1994): A comparison of order estimation procedures for ARMA models. *Statistica Sinica* 4, 29-50.

Quinn, B. G. (1980): Order determination for a multivariate autoregression. *Journal of the Royal Statistical Society, Series B* 42, 182-185.

Quinn, B. G. (1988): A note on AIC order determination for multivariate autoregressions. *Journal of Time Series Analysis* 9, 241-245.

Rao, C. R. & Y. Wu (1989): A strongly consistent procedure for model selection in a regression problem. *Biometrika* 76, 369-374.

Rao, C. R. & Y. Wu (2001): On model selection. In: P. Lahiri (ed.), Model Selection. *IMS Lecture Notes Monograph Series* Vol 38, 1–57.

Rao, C. R. & Y. Wu, Y. (2005): Linear model selection by cross-validation. *Journal of Statistical Planning and Inference* 128, 231–240.

Reschenhofer, E. (1999): Improved estimation of the expected Kullback-Leibler discrepancy in case of misspecification. *Econometric Theory* 15, 377-387.

Rissanen, J. (1978): Modeling by shortest data description. *Automatica* 14, 465–471.

Rissanen, J. (1983): A universal prior for integers and estimation by minimum description length. *Annals of Statistics* 11, 416–431.

Rissanen, J. (1986a): Stochastic complexity and modeling. *Annals of Statistics* 14, 1080–1100.

Rissanen, J. (1986b): A predictive least squares principle. *IMA Journal of Mathematical Control and Information* 3, 211–222.

Rissanen, J. (1987): Stochastic complexity (with discussion). *Journal of the Royal Statistical Society, Series B* 49, 223–265.

Rissanen, J. (1989): *Stochastic Complexity and Statistical Inquiry*. Teaneck: World Scientific.

Sakai, H. (1981): Asymptotic distribution of the order selected by AIC in multivariate autoregressive model fitting. *International Journal of Control* 33, 175-180.

Saleh, A. K. M. E. (2006): *Theory of Preliminary Test and Stein-Type Estimation with Applications*. Hoboken: Wiley.

Saleh, A. K. M. E. & P. K. Sen (1983): Asymptotic properties of tests of hypothesis following a preliminary test. *Statistics and Decisions* 1, 455–477.

Sawa, T. & T. Hiromatsu (1973): Minimax regret significance points for a preliminary test in regression analysis. *Econometrica* 41, 1093-1101.

Schwarz, G. (1978): Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.

Sclove, S. L., Morris, C. & R. Radhakrishnan (1972): Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Annals of Mathematical Statistics* 43, 1481-1490.

Sen, P. K (1979): Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics* 7, 1019–1033.

Sen, P. K & A. K. M. E. Saleh (1987): On preliminary test and shrinkage M-estimation in linear models. *Annals of Statistics* 15, 1580–1592.

Shao, J. (1993): Linear model selection by cross-validation. *Journal of the American Statistical Association* 88, 486-494.

Shao, J. (1997): An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* 7, 221-264.

Shao, J. (1998): Convergence rates of the generalized information criterion. *Journal of Nonparametric Statistics* 9, 217-225.

Shen, X., Huang, H. C. & J. Ye (2004): Inference after model selection. *Journal of the American Statistical Association* 99, 751-762.

Shibata, R. (1976): Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* 63, 117-126.

Shibata, R. (1980): Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8, 147–164.

Shibata, R. (1981a): An optimal autoregressive spectral estimate. *Annals of Statistics* 9, 300-306.

Shibata, R. (1981b): An optimal selection of regression variables. *Biometrika* 68, 45–54. (Correction: *Biometrika* 69 (1982), 492.)

Shibata, R. (1984): Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* 71, 43-49.

Shibata, R. (1986a): Consistency of model selection and parameter estimation. *Journal of Applied Probability, Special Volume* 23A, 127-141.

Shibata, R. (1986b): Selection of the number of regression variables; a minimax choice of generalized FPE. *Annals of the Institute of Statistical Mathematics* 38, 459–474.

Shibata, R. (1989): Statistical aspects of model selection. In: J. C. Willems (ed.), *From Data to Model*, 215-240. Springer-Verlag.

Shibata, R. (1997): Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica* 7, 375-394.

Söderström, T. (1977): On model structure testing in system identification. *International Journal of Control* 26, 1-18.

Stone, C. (1981): Admissible selection of an accurate and parsimonious normal linear regression model. *Annals of Statistics* 9, 475-485.

Stone, C. (1982): Local asymptotic admissibility of a generalization of Akaike's model selection rule. *Annals of the Institute of Statistical Mathematics* 34, 123-133.

Stone, M. (1974): Cross-validatory choice and assessment of statistical prediction. *Journal of the Royal Statistical Society, Series B* 36, 111-133.

Stone, M. (1977): An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* 39, 44-47.

Strawderman, W. E. (1971): Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics* 42, 385–388.

Sugiura, N. (1978): Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics* A 7, 13-26.

Takada, Y. (1982): Admissibility of some variable selection rules in linear regression model. *Journal of the Japanese Statistical Society* 12, 45-49.

Takeuchi, K. (1976): Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku* 153, 12-18. (In Japanese.)

Teräsvirta, T. & I. Mellin (1986): Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics* 13, 159-171.

Theil, H. (1961): *Economic Forecasts and Policy.* 2nd edition. Amsterdam: North-Holland.

Thompson, M. L. (1978a): Selection of variables in multiple regression: part I. A review and evaluation. *International Statistical Review* 46, 1-19.

Thompson, M. L. (1978b): Selection of variables in multiple regression: part II. Chosen procedures, computations and examples. *International Statistical Review* 46, 129-146.

Tibshirani, R. (1996): Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* 58, 267-288.

Toro-Vizcarrondo, C. & T. D. Wallace (1968) A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association* 63, 558–572.

Toyoda, T. & T. D. Wallace (1976): Optimal critical values for pre-testing in regression. *Econometrica* 44, 365-375.

Tsay, R. S. (1984): Order selection in nonstationary autoregressive models. *Annals of Statistics* 12, 1425-1433.

Venter, J. H. & S. J. Steele (1992): Some contributions to selection and estimation in the normal linear model. *Annals of the Institute of Statistical Mathematics* 44, 281-297.

Vuong, Q. H. (1989): Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307-333.

Wallace, T. D. (1972): Weaker criteria and tests for linear restrictions in regression. *Econometrica* 40, 689–698.

Wei, C. Z. (1992): On predictive least squares principles. *Annals of Statistics* 20, 1-42.

Wegkamp, M. (2003): Model selection in nonparametric regression. *Annals of Statistics* 31, 252–273.

Yang, Y. (1999): Model selection for nonparametric regression. *Statistica Sinica* 9, 475–499.

Yang, Y. (2001): Adaptive regression by mixing. *Journal of the American Statistical Association* 96, 574–588.

Yang, Y. (2003): Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica* 13, 783–809.

Yang, Y. (2005): Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92, 937–950.

Yang, Y. (2007): Prediction/estimation with simple linear models: Is it really that simple? *Econometric Theory* 23, 1–36.

Yang, Y. & A. R. Barron (1998): An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory* 44, 95–116.

Yang, Y. & A. R. Barron (1999): Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* 27, 1564-1599.

Zhang, P. (1992): Inference after variable selection in linear regression models. *Biometrika* 79, 741–746.

Zhang, P. (1993a): Model selection via multifold cross validation. *Annals of Statistics* 21, 299-313.

Zhang, P. (1993b): On the convergence rate of model selection criteria. *Communications in Statistics. Theory and Methods* 22, 2765-2775.

Zheng, X. & W. Y. Loh (1995): Consistent variable selection in linear models. *Journal of the American Statistical Association* 90 151–156.

Zheng, X. & W. Y. Loh (1997): A consistent variable selection criterion for linear models with high-dimensional covariates. *Statistica Sinica* 7, 311–325.

Zou, H. (2006): The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101, 1418–1429.