4

A comparison of some selection methods

In this chapter we compare some information criteria with respect to consistency and efficiency, which are classical themes in model selection. The comparison is driven by a study of the 'penalty' applied to the maximised log-likelihood value, in a framework with increasing sample size. AIC is not strongly consistent, though it is efficient, while the opposite is true for the BIC. We also introduce Hannan and Quinn's criterion, which has properties similar to those of the BIC, while Mallows's C_p and Akaike's FPE behave like AIC.

4.1 Comparing selectors: consistency, efficiency and parsimony

If we make the assumption that there exists one true model that generated the data and that this model is one of the candidate models, we would want the model selection method to identify this true model. This is related to consistency. A model selection method is weakly consistent if, with probability tending to one as the sample size tends to infinity, the selection method is able to select the true model from the candidate models. Strong consistency is obtained when the selection of the true model happens almost surely. Often, we do not wish to make the assumption that the true model is amongst the candidate models. If instead we are willing to assume that there is a candidate model that is closest in Kullback–Leibler distance to the true model, we can state weak consistency as the property that, with probability tending to one, the model selection method picks such a closest model. In this chapter we explain why some information criteria have this property, and others do not. A different nice property that we might want an information criterion to possess is that it behaves 'almost as well', in terms of mean squared error, or expected squared prediction error, as the theoretically best model for the chosen type of squared error loss. Such a model selection method is called efficient. In Section 4.5 we give a more precise statement of the efficiency of an information criterion, and identify several information criteria that are efficient.

At the end of the chapter we will explain that consistency and efficiency cannot occur together, thus, in particular, a consistent criterion can never be efficient.

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

Several information criteria have a common form, which is illustrated by AIC, see (2.1), and the BIC, see (3.1):

$$\begin{aligned}\mathrm{AIC}\{f(\cdot;\theta)\} &= 2\ell_n(\widehat{\theta}) - 2\,\mathrm{length}(\theta),\\ \mathrm{BIC}\{f(\cdot;\theta)\} &= 2\ell_n(\widehat{\theta}) - (\log n)\,\mathrm{length}(\theta).\end{aligned}$$

Both criteria are constructed as twice the maximised log-likelihood value minus a penalty for the complexity of the model. BIC's penalty is larger than that of AIC, for all n at least 8. This shows that the BIC more strongly discourages choosing models with many parameters.

For the kth model in our selection list ($k = 1, \ldots, K$), denote the parameter vector by θ_k and the density function for the ith observation by $f_{k,i}$. For the regression situation, $f_{k,i}(y_i, \theta_k) = f_k(y_i \mid x_i, \theta_k)$. This density function depends on the value of the covariate x_i for the ith observation. For the i.i.d. case there is no such dependence and $f_{k,i} = f_k$, for all observations. In its most general form, the data are not assumed to be independent.

Both information criteria AIC and BIC take the form

$$\mathrm{IC}(M_k) = 2\sum_{i=1}^{n} \log f_{k,i}(Y_i, \widehat{\theta}_k) - c_{n,k},$$

where $c_{n,k} > 0$ is the penalty for model M_k, for example, $c_{n,k} = 2\,\mathrm{length}(\theta_k)$ for AIC. The better model has the larger value of IC. The factor 2 is not really needed, though it is included here for historical reasons. Other examples of criteria which take this form are AIC_c, TIC and BIC^*.

A parsimonious model. One underlying purpose of model selection is to use the information criterion to select the model that is closest to the (unknown) but true model. The true data-generating density is denoted by $g(\cdot)$. The Kullback–Leibler distance, see equation (2.2), can be used to measure the distance from the true density to the model density. If there are two or more models that minimise the Kullback–Leibler distance, we wish to select that model which has the fewest parameters. This is called the most parsimonious model.

Sin and White (1996) give a general treatment on asymptotic properties of information criteria of the form $\mathrm{IC}(M_k)$. We refer to that paper for precise assumptions on the models and for proofs of the results phrased in this section.

Theorem 4.1 ***Weak consistency.*** *Suppose that amongst the models under consideration there is exactly one model M_{k_0} which reaches the minimum Kullback–Leibler distance. That is, for this model it holds that*

$$\liminf_{n\to\infty} \min_{k\neq k_0} n^{-1} \sum_{i=1}^{n} \{\mathrm{KL}(g, f_{k,i}) - \mathrm{KL}(g, f_{k_0,i})\} > 0.$$

Let the strictly positive penalty be such that $c_{n,k} = o_p(n)$. Then, with probability going to one, the information criterion selects this closest model M_{k_0} as the best model.

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

Thus, in order to pick the Kullback–Leibler best model with probability going to one, or in other words, to have weak consistency of the criterion, the condition on the penalty is that when it is divided by n, it should tend to zero for growing sample size. Note that for the result on weak consistency to hold it is possible to have one of the $c_{n,k} = 0$, while all others are strictly positive.

As a consequence of the theorem, we immediately obtain the weak consistency of the BIC where $c_{n,k} = (\log n)\,\text{length}(\theta_k)$. The fixed penalty $c_{n,k} = 2\,\text{length}(\theta_k)$ of AIC also satisfies the condition, and hence AIC is weakly consistent under these assumptions.

Suppose now that amongst the models under consideration there are two or more models which reach the minimum Kullback–Leibler distance. Parsimony means that amongst these models the model with the fewest number of parameters, that is, the 'simplest model', is chosen. In the literature, this parsimony property is sometimes referred to as consistency, hereby ignoring the situation where there is a unique closest model. Consistency can be obtained under two types of technical condition. Denote by $\mathcal{J}$ the set of indices of the models which all reach the miminum Kullback–Leibler distance to the true model, and denote by $\mathcal{J}_0$ the subset of $\mathcal{J}$ containing models with the smallest dimension (there can be more than one such smallest model).

Theorem 4.2 ***Consistency.*** *Assume either set of conditions (a) or (b).*

(a) Assume that for all $k_0 \neq \ell_0 \in \mathcal{J}$*:*

$$\limsup_{n\to\infty} n^{-1/2} \sum_{i=1}^{n} \{\mathrm{KL}(g, f_{k_0,i}) - \mathrm{KL}(g, f_{\ell_0,i})\} < \infty.$$

For any index j_0 *in* $\mathcal{J}_0$*, and for any index* $\ell \in \mathcal{J}\backslash\mathcal{J}_0$*, let the penalty be such that* $\mathrm{P}\{(c_{n,\ell} - c_{n,j_0})/\sqrt{n} \to \infty\} = 1$.

(b) Assume that for all $k_0 \neq \ell_0 \in \mathcal{J}$*, the log-likelihood ratio*

$$\sum_{i=1}^{n} \log \frac{f_{k_0,i}(Y_i, \theta^*_{k_0})}{f_{\ell_0,i}(Y_i, \theta^*_{\ell_0})} = O_p(1),$$

and that for any j_0 *in* $\mathcal{J}_0$ *and* $\ell \in \mathcal{J}\backslash\mathcal{J}_0$*,* $\mathrm{P}(c_{n,\ell} - c_{n,j_0} \to \infty) = 1$.

Then, with probability tending to one the information criterion will pick such a smallest model:

$$\lim_{n\to\infty} \mathrm{P}\left\{ \min_{\ell\in\mathcal{J}\backslash\mathcal{J}_0} \big(\mathrm{IC}(M_{j_0}) - \mathrm{IC}(M_\ell)\big) > 0 \right\} = 1.$$

Part (b) requires boundedness in distribution of the log-likelihood ratio statistic. The asymptotic distribution of such statistic is studied in large generality by Vuong (1989). The most well-known situation is when the models are nested. In this case it is known that twice the maximised log-likelihood ratio value follows asymptotically a chi-squared distribution with degrees of freedom equal to the difference in number of parameters of the two models, when the smallest model is true. Since the limiting distribution is a

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

chi-squared distribution, this implies that it is bounded in probability (that is, $O_P(1)$), and hence the condition in (b) is satisfied.

The BIC penalty $c_{n,k} = (\log n)\text{length}(\theta_k)$ obviously satisfies the penalty assumption in (b), but the AIC penalty fails. Likewise, the AIC_c and TIC penalties fail this assumption.

In fact, criteria with a fixed penalty, not depending on sample size, do not satisfy either penalty condition in Theorem 4.2. This implies that, for example, AIC will not necessarily choose the most parsimonious model, there is a probability of overfitting. This means that the criterion might pick a model that has more parameters than actually needed. Hence with such criteria, for which the assumptions of Theorem 4.2 do not hold, there is a probability of selecting too many parameters when there are two or more models which have minimal Kullback–Leibler distance to the true model. We return to the topic of overfitting in Section 8.3.

4.2 Prototype example: choosing between two normal models

A rather simple special case of the general model selection framework is the following. Observations $Y_1, \ldots, Y_n$ are i.i.d. from the normal density $\text{N}(\mu, 1)$, with two models considered: model M_0 assumes that $\mu = 0$, while model M_1 remains 'general' and simply says that $\mu \in \mathbb{R}$. We investigate consequences of using different penalty parameters $c_{n,k} = d_n \,\text{length}(\theta_k)$. The values of the information criteria are

$$\begin{aligned} \text{IC}_0 &= 2 \max_{\mu} \{\ell_n(\mu) \colon M_0\} - d_n \cdot 0, \\ \text{IC}_1 &= 2 \max_{\mu} \{\ell_n(\mu) \colon M_1\} - d_n \cdot 1. \end{aligned}$$

The log-likelihood function here is that of a normal model with unknown mean μ and known variance equal to one,

$$\ell_n(\mu) = -\tfrac{1}{2} \sum_{i=1}^{n} (Y_i - \mu)^2 - \tfrac{1}{2} n \log(2\pi) = -\tfrac{1}{2} n (\bar{Y} - \mu)^2 - \tfrac{1}{2} n \{\widehat{\sigma}^2 + \log(2\pi)\}.$$

The right-hand side of this equation is obtained by writing $Y_i - \mu = Y_i - \bar{Y} + \bar{Y} - \mu$, where $\bar{Y}$ as usual denotes the average and $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. Apart from the additive terms $-\frac{1}{2} n (\widehat{\sigma}^2 + \log(2\pi))$, not depending on the models,

$$\text{IC}_0 = -\tfrac{1}{2} n \bar{Y}^2 \quad \text{and} \quad \text{IC}_1 = -d_n,$$

showing that

$$\text{selected model} = \begin{cases} M_1 & \text{if } |\sqrt{n}\bar{Y}| \geq d_n^{1/2}, \\ M_0 & \text{if } |\sqrt{n}\bar{Y}| < d_n^{1/2}. \end{cases} \tag{4.1}$$

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

Secondly, the resulting estimator of the mean parameter is

$$\widehat{\mu} = \begin{cases} \bar{Y} & \text{if } |\sqrt{n}\bar{Y}| \geq d_n^{1/2}, \\ 0 & \text{if } |\sqrt{n}\bar{Y}| < d_n^{1/2}. \end{cases} \tag{4.2}$$

Primary candidates for the penalty factor include $d_n = 2$ for AIC and $d_n = \log n$ for the BIC. We now investigate consequences of such and similar choices, in the large-sample framework where n increases.

We first apply Theorem 4.1. Under the assumption that the biggest model is the true model, the true density g equals the $\mathrm{N}(\mu, 1)$ density. Obviously, for model M_1 the Kullback–Leibler distance to the true model is equal to zero. For model M_0, this distance is equal to $\frac{1}{2}\mu^2$, which shows that the assumption of Theorem 4.1 on the KL-distances holds. In particular, both AIC and the BIC will consistently select the wide model, the model with the most variables, as the best one. However, if the true model is model M_0, then $\mu = 0$ and the difference in Kullback–Leibler values equals zero. Both models have the same Kullback–Leibler distance. For such a situation Theorem 4.1 is not applicable, instead we use Theorem 4.2. Since the two models with distributions $\mathrm{N}(0, 1)$ and $\mathrm{N}(\mu, 1)$ are nested, the limit of the log-likelihood ratio statistic, with μ estimated by its maximum likelihood estimator, has a χ_1^2 distribution if the $\mathrm{N}(0, 1)$ model is true, and hence we need the additional requirement that the penalty diverges to infinity for growing sample size n in order to select the most parsimonious model with probability one. Only the BIC leads to consistent model selection in this case, AIC does not.

The model selection probabilities for the AIC scheme are easily found here. With Z standard normal, they are given by

$$\mathrm{P}_n(M_1 \mid \mu) = \mathrm{P}(|\sqrt{n}\mu + Z| \leq \sqrt{2}).$$

Figure 4.1 shows the probability $\mathrm{P}_n(M_1 \mid \mu)$ for the AIC and BIC schemes, for a high number of data points, $n = 1000$. It illustrates that the practical difference between the two methods might not be very big, and that also AIC has detection probabilities of basically the same shape as with the BIC. We note that $\mathrm{P}_n(M_1 \mid 0) = \mathrm{P}(\chi_1^2 \geq d_n)$, which is 0.157 for the AIC method and which goes to zero for the BIC.

Next consider the risk performance of $\widehat{\mu}$, which we take to be the mean squared error multiplied by the sample size; this scaling is natural since variances of regular estimators are $O(1/n)$. We find

$$r_n(\mu) = n\,\mathrm{E}_\mu\{(\widehat{\mu} - \mu)^2\} = \mathrm{E}\big[\big\{(\sqrt{n}\mu + Z)I\{|\sqrt{n}\mu + Z| \geq d_n^{1/2}\} - \sqrt{n}\mu\big\}^2\big],$$

which can be computed and plotted via numerical integration or via an explicit formula. In fact, as seen via Exercise 4.3,

$$r_n(\mu) = 1 - \int_{\mathrm{low}_n}^{\mathrm{up}_n} z^2\phi(z)\,\mathrm{d}z + n\mu^2\{\Phi(\mathrm{up}_n) - \Phi(\mathrm{low}_n)\}, \tag{4.3}$$

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

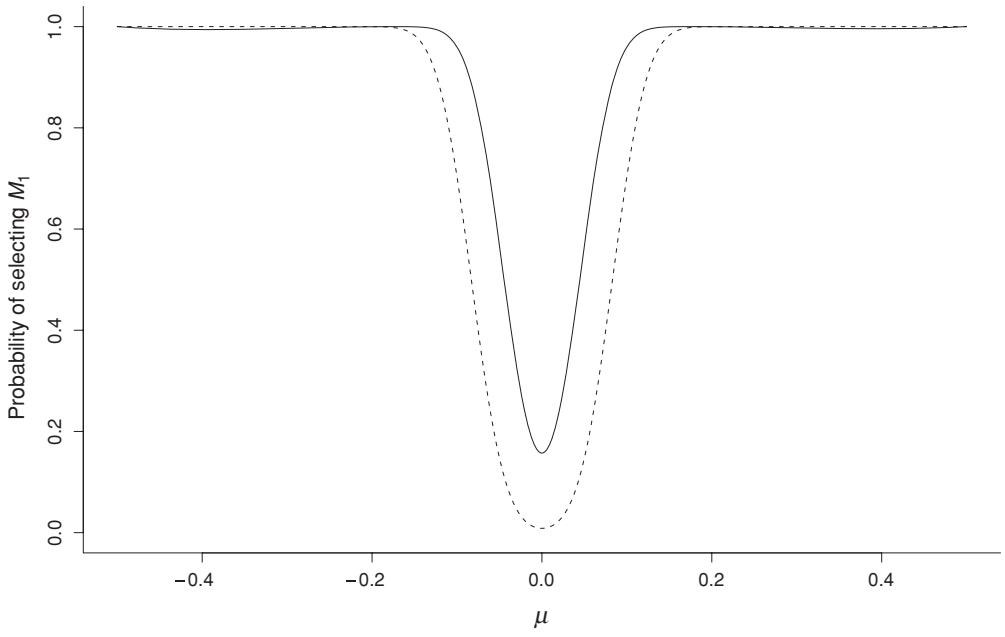

Fig. 4.1. The probability that model M_1 is selected, for the AIC (full line) and the BIC method (dashed line), for $n = 1000$ data points.

where

$$\text{low}_n = -d_n^{1/2} - \sqrt{n}\mu \quad \text{and} \quad \text{up}_n = d_n^{1/2} - \sqrt{n}\mu.$$

Risk functions are plotted in Figure 4.2 for three information criteria: the AIC with $d_n = 2$, the BIC with $d_n = \log n$, and the scheme that uses $d_n = \sqrt{n}$. Here AIC does rather better than the BIC; in particular, its risk function is bounded (with maximal value 1.647, for all n, actually), whereas $r_n(\mu)$ for the BIC exhibits much higher maximal risk. We show in fact below that its max-risk $r_n = \max r_n(\mu)$ is unbounded, diverging to infinity as n increases.

The information criterion with $d_n = \sqrt{n}$ is included in Figure 4.2, and in our discussion, since it corresponds to the somewhat famous estimator

$$\widehat{\mu}_H = \begin{cases} \bar{Y} & \text{if } |\bar{Y}| \geq n^{-1/4}, \\ 0 & \text{if } |\bar{Y}| < n^{-1/4}. \end{cases}$$

The importance of this estimator, which stems from J. L. Hodges Jr., cf. Le Cam (1953), is not that it is meant for practical use, but that it exhibits what is known as 'superefficiency' at zero:

$$\sqrt{n}(\widehat{\mu}_H - \mu) \xrightarrow{d} \begin{cases} \text{N}(0, 1) & \text{for } \mu \neq 0, \\ \text{N}(0, 0) & \text{for } \mu = 0. \end{cases} \tag{4.4}$$

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

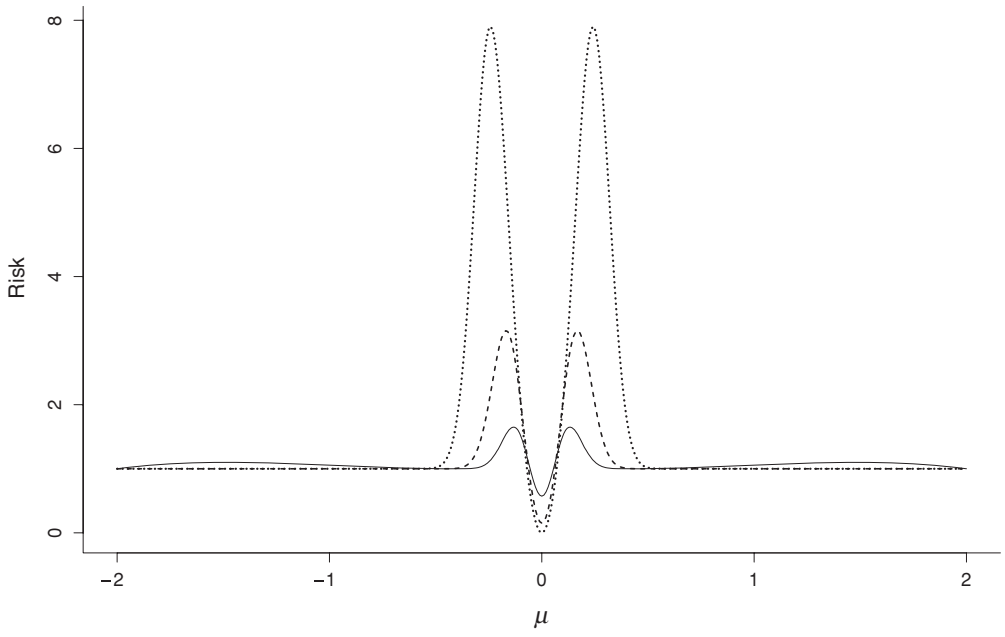

Fig. 4.2. Risk functions for estimators $\widehat{\mu}$ of 4.2, for $n = 200$, for penalties d_n equal to 2 (AIC, full line), $\log n$ (BIC, dashed line), $\sqrt{n}$ (Hodges, dotted line), with maximal risks equal to 1.647, 3.155, 7.901 respectively. The conservative estimator $\widehat{\mu} = \bar{Y}$ has constant risk function equal to 1.

Here the limit constant 0 is written as N(0, 0), to emphasise more clearly the surprising result that the variance of the limit distribution is $v(\mu) = 1$ for $\mu \neq 0$ but $v(\mu) = 0$ for $\mu = 0$. The result is not only surprising but also alarming in that it appears to clash with the Cramér–Rao inequality and with results about the asymptotic optimality of maximum likelihood estimators. At the same time, there is the unpleasant behaviour of the risk function close to zero for large n, as illustrated in Figure 4.2. Thus there are crucial differences between pointwise and uniform approximations of the risk functions; result (4.4) is mathematically true, but only in a restricted pointwise sense.

It can be shown that consistency at the null model (an attractive property) actually implies unlimited max-risk with growing n (an unattractive property). To see this, consider the risk expression (4.3) at points $\mu = \delta/\sqrt{n}$, where

$$r_n(\delta/\sqrt{n}) \geq \delta^2\{\Phi(\delta + d_n^{1/2}) - \Phi(\delta - d_n^{1/2})\}.$$

At the particular point $\mu = (d_n/n)^{1/2}$, therefore, the risk is bigger than the value $d_n\{\Phi(2d_n^{1/2}) - \frac{1}{2}\}$, which diverges to infinity precisely when $d_n \to \infty$, which happens if and only if $\mathrm{P}_n(M_0 \mid \mu = 0) \to 1$. That is, when there is consistency at the null model. As illustrated in Figure 4.2, the phenomenon is more pronounced for the Hodges estimator than for the BIC. The aspect that an estimator-post-BIC carries max risk proportional

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

to $\log n$, and that this max risk is met close to the null model at which consistency is achieved, remains however a somewhat negative property.

We have uncovered a couple of crucial differences between the AIC and BIC methods. (i) The BIC behaves very well from the consistency point of view; with large n it gives a precise indication of which model is correct. This might or might not be relevant in the context in which it is used. (ii) But it pays a price for its null model consistency property: the risk function for the μ estimator exhibits unpleasant behaviour near the null model, and its maximal risk is unbounded with increasing sample size. In this respect AIC fares rather better.

4.3 Strong consistency and the Hannan–Quinn criterion

Weak consistency is defined via weak convergence: it states that with probability converging to one, a most parsimonious model with minimum Kullback–Leibler distance to the true model is selected. The results can be made stronger by showing that under some conditions such a model is selected almost surely. This is called strong consistency.

Theorem 4.3 ***Strong consistency.*** *Suppose that amongst the models under consideration there is exactly one model M_{k_0} which reaches the minimum Kullback–Leibler distance, such that*

$$\liminf_{n\to\infty} \min_{k\neq k_0} n^{-1} \sum_{i=1}^{n} \{\mathrm{KL}(g, f_{k,i}) - \mathrm{KL}(g, f_{k_0,i})\} > 0.$$

Let the strictly positive penalty be such that $c_{n,k} = o(n)$ almost surely. Then

$$\mathrm{P}\Big\{ \min_{\ell\neq k_0} \big(\mathrm{IC}(M_{k_0}) - \mathrm{IC}(M_\ell)\big) > 0, \text{ for almost all } n \Big\} = 1.$$

The assumption on the penalty is very weak, and is satisfied for the nonstochastic penalties of AIC and the BIC. This tells us that there are situations where AIC, and the BIC, are strongly consistent selectors of the model that is best in minimising the Kullback–Leibler distance to the true model.

The situation is more complex if there is more than one model reaching minimum Kullback–Leibler distance to the true model. A general treatment of this subject is presented in Sin and White (1996). For details we refer to that paper. The main difference between showing weak and strong consistency is, for the latter, the use of the law of the iterated logarithm instead of the central limit theorem.

Using the notation of Theorem 4.2, if for all $k_0 \neq \ell_0 \in \mathcal{J}$:

$$\limsup_{n\to\infty} \frac{1}{\sqrt{n \log\log n}} \sum_{i=1}^{n} \{\mathrm{KL}(g, f_{k_0,i}) - \mathrm{KL}(g, f_{\ell_0,i})\} \leq 0,$$

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

then (under some additional assumptions) the requirement on the penalty $c_{n,k}$ to guarantee strong consistency of selection is that

$$P(c_{n,k} \geq a_n \sqrt{n \log\log n} \text{ for almost all } n) = 1,$$

where a_n is a random sequence, almost surely bounded below by a strictly positive number.

If it is rather the situation that for all $k_0 \neq \ell_0 \in \mathcal{J}$, the log-likelihood ratio

$$\sum_{i=1}^{n} \log \frac{f_{k_0,i}(Y_i, \theta^*_{k_0})}{f_{\ell_0,i}(Y_i, \theta^*_{\ell_0})} = o(\log\log n) \text{ almost surely},$$

then the required condition on the penalty is that

$$\mathrm{P}(c_{n,k} \geq b_n \log\log n \text{ for almost all } n) = 1,$$

where b_n is a random sequence, almost surely bounded below by a strictly positive number.

For a sequence of strictly nested models, the second condition on the penalty is sufficient (Sin and White, 1996, Corollary 5.3).

The BIC penalty $c_{n,k} = \log n \, \text{length}(\theta_k)$ satisfies these assumptions, leading to strong consistency of model selection, provided the other assumptions hold. The above results indicate that $\log n$ is not the slowest rate by which the penalty can grow to infinity in order to almost surely select the most parsimonious model. The application of the law of the iterated logarithm to ensure strong consistency of selection leads to Hannan and Quinn's (1979) criterion

$$\mathrm{HQ}\{f(\cdot;\theta)\} = 2 \log \mathcal{L}(\widehat{\theta}) - 2c \log\log n \, \text{length}(\theta), \text{ with } c > 1.$$

The criterion was originally derived to determine the order in an autoregressive time series model. Hannan and Quinn (1979) do not give any advice on which value of c to choose. Note that for practical purposes, this choice of penalty might not be that useful. Indeed, even for very large sample sizes the quantity $\log\log n$ remains small, and whatever method is used for determining a threshold value c will be more important than the $\log\log n$ factor in determining the value of $2c \log\log n$.

4.4 Mallows's C_p and its outlier-robust versions

A criterion with behaviour similar to that of AIC for variable selection in regression models is Mallows's (1973)

$$C_p = \mathrm{SSE}_p / \widehat{\sigma}^2 - (n - 2p). \tag{4.5}$$

Here SSE_p is the residual sum of squares $\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$ in the model with p regression coefficients, and the variance is computed in the largest model. C_p is defined as an

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

estimator of the scaled squared prediction error

$$\mathrm{E}\left\{\sum_{i=1}^{n}(\widehat{Y}_i - \mathrm{E}Y_i)^2\right\}\Big/\sigma^2.$$

If the model with p variables contains no bias, then C_p is close to p. If, on the other hand, there is a large bias, then $C_p > p$. Values close to the corresponding p (but preferably smaller than p) indicate a good model. A plot of C_p versus p often helps in identifying good models.

Ronchetti and Staudte (1994) define an outlier-robust version of Mallows's C_p, as an estimator of $\mathrm{E}\{\sum_{i=1}^{n}\widehat{w}_i^2(\widehat{Y}_i - \mathrm{E}Y_i)^2\}/\sigma^2$, which is a weighted scaled squared prediction error, using the robust weights defined in Section 2.10.3. In detail,

$$RC_p = W_p\widehat{\sigma}^2 - (U_p - V_p),$$

where we need the following definitions. With $\eta(x, \varepsilon)$ as defined in Section 2.10.3, define $M = \mathrm{E}\{(\partial/\partial\varepsilon)\eta(x, \varepsilon)xx^{\mathrm{t}}\}$, $W = \mathrm{E}(w^2xx^{\mathrm{t}})$, $R = \mathrm{E}\{\eta^2(x, \varepsilon)xx^{\mathrm{t}}\}$, $N = \mathrm{E}\{\eta^2(\partial/\partial\varepsilon)\eta xx^{\mathrm{t}}\}$ and $L = \mathrm{E}[\{((\partial/\partial\varepsilon)\eta)^2 + 2(\partial/\partial\varepsilon)\eta w - 3w^2\}xx^{\mathrm{t}}]$. Then we let $V_p = \mathrm{Tr}(WM^{-1}RM^{-1})$ and

$$U_p - V_p = \sum_{i=1}^{n}\mathrm{E}\{\eta^2(x_i, \varepsilon_i)\} - 2\,\mathrm{Tr}(NM^{-1}) + \mathrm{Tr}(LM^{-1}RM^{-1}).$$

In practice, approximations to U_p and V_p are used, see Ronchetti and Staudte (1994). The robust C_p is, for example, available in the S-Plus library `Robust`.

A different robustification of C_p is arrived at via a weighting scheme similar to that used to obtain the weighted AIC in Section 2.10.1. The weight functions $w(\cdot)$ are here based on smoothed Pearson residuals. This leads to

$$WC_p = \widehat{\sigma}^{-2}\sum_{i=1}^{n} w(y_i - x_i^{\mathrm{t}}\widehat{\beta}, \widehat{\sigma})(y_i - x_i^{\mathrm{t}}\widehat{\beta})^2 - \sum_{i=1}^{n} w(y_i - x_i^{\mathrm{t}}\widehat{\beta}, \widehat{\sigma}) + 2\,p.$$

Agostinelli (2002) shows that without the presence of outliers the WC_p is asymptotically equivalent to C_p. When the weight function $w(\cdot) \equiv 1$, then $WC_p = C_p$.

4.5 Efficiency of a criterion

Judging model selection criteria is not only possible via a study of the selection of the most parsimonious subset of variables (consistency issues). The approach taken in this section is a study of the efficiency of the criterion with respect to a loss function.

Let us consider an example. We wish to select the best set of variables in the regression model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \varepsilon_i, \quad i = 1, \ldots, n,$$

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

where $\mathrm{Var}\,\varepsilon_i = \sigma^2$, with the specific purpose of predicting a new (independent) outcome variable $\widehat{Y}_i$ at the observed covariate combination $x_i = (x_{i,1}, \ldots, x_{i,k})^{\mathrm{t}}$, for $i = 1, \ldots, n$. For prediction, the loss is usually taken to be the squared prediction error. This means that we wish to select that set S of covariates x_j, $j \in S$ for which the expected prediction error, conditional on the observed data $\mathcal{Y}_n = (Y_1, \ldots, Y_n)$, is as small as possible. That is, we try to minimise

$$\sum_{i=1}^{n} \mathrm{E}\{(\widehat{Y}_{S,i} - Y_{\mathrm{true},i})^2 | \mathcal{Y}_n\}, \tag{4.6}$$

where it is understood that the $Y_{\mathrm{true},i}$ are independent of $Y_1, \ldots, Y_n$, but otherwise come from the same distribution. The predicted values $\widehat{Y}_{S,i}$ depend on the data $\mathcal{Y}$. The notation $\widehat{Y}_{S,i}$ indicates that the prediction is made using the model containing only the covariates x_j, $j \in S$. Using the independence between the new observations and the original $\mathcal{Y}_n = (Y_1, \ldots, Y_n)$, we can write

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{E}\{(\widehat{Y}_{S,i} - Y_{\mathrm{true},i})^2 | \mathcal{Y}_n\} &= \sum_{i=1}^{n} \mathrm{E}[\{\widehat{Y}_{S,i} - \mathrm{E}(Y_{\mathrm{true},i})\}^2 | \mathcal{Y}_n] + n\sigma^2 \\ &= \sum_{i=1}^{n} (\widehat{\beta}_S - \beta_{\mathrm{true}})^{\mathrm{t}} x_i x_i^{\mathrm{t}} (\widehat{\beta}_S - \beta_{\mathrm{true}}) + n\sigma^2.\end{aligned}$$

To keep the notation simple, we use β_S also for the vector with zeros inserted for undefined entries, to make the dimensions of β_S and β_{true} equal. If the model selected via a model selection criterion reaches the minimum of (4.6) as the sample size n tends to infinity, we call the selection criterion *efficient*, conditional on the given set of data. Instead of conditioning on the given sample $\mathcal{Y}_n$, theoretically, we rather work with the (unconditional) expected prediction error

$$L_n(S) = \sum_{i=1}^{n} \mathrm{E}\{(\widehat{Y}_{S,i} - Y_{\mathrm{true},i})^2\}, \tag{4.7}$$

which for the regression situation reads

$$L_n(S) = \mathrm{E}\{(\widehat{\beta}_S - \beta_{\mathrm{true}})^{\mathrm{t}} X^{\mathrm{t}} X (\widehat{\beta}_S - \beta_{\mathrm{true}})\} + n\sigma^2. \tag{4.8}$$

Denote by S^* the index set for which the minimum value of the expected prediction error is attained. Let $\widehat{S}$ be the set of indices in the selected model. The notation $\mathrm{E}_{\widehat{S}}$ denotes that the expectation is taken with respect to all random quantities except for $\widehat{S}$. The criterion used to select $\widehat{S}$ is efficient when

$$\frac{\sum_{i=1}^{n} \mathrm{E}_{\widehat{S}}\{(\widehat{Y}_{\widehat{S},i} - Y_{\mathrm{true},i})^2\}}{\sum_{i=1}^{n} \mathrm{E}\{(\widehat{Y}_{S^*,i} - Y_{\mathrm{true},i})^2\}} = \frac{L_n(\widehat{S})}{L_n(S^*)} \xrightarrow{p} 1, \quad \text{as } n \to \infty.$$

In other words, a model selection criterion is called *efficient* if it selects a model such that the ratio of the expected loss function at the selected model and the expected loss

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

function at its theoretical minimiser converges to one in probability. Note that since we use a selected model $\widehat{S}$, the numerator is a random variable.

When focus is on estimation rather than prediction, we look at the squared estimation error. After taking expectations this leads to studying the mean squared error. We show below, in the context of autoregressive models and normal linear regression, that some model selectors, including AIC, are efficient.

4.6 Efficient order selection in an autoregressive process and the FPE

Historically, the efficiency studies originate from the time series context where a model is sought that asymptotically minimises the mean squared prediction error. Shibata (1980) formulates the problem for the order selection of a linear time series model. Let ε_i $(i = \ldots, -1, 0, 1, \ldots)$ be independent and identically distributed $\mathrm{N}(0, \sigma^2)$ random variables. Consider the stationary Gaussian process $\{X_i\}$ such that for real-valued coefficients a_j,

$$X_i - (a_1 X_{i-1} + a_2 X_{i-2} + \cdots) = \varepsilon_i.$$

In practice, the time series is truncated at order k, resulting in the kth-order autoregressive model, built from observations $X_1, \ldots, X_n$,

$$X_i - (a_1 X_{i-1} + \cdots + a_k X_{i-k}) = \varepsilon_i, \quad i = 1, \ldots, n. \tag{4.9}$$

In this model we estimate the unknown coefficients $\boldsymbol{a}_k = (a_1, \ldots, a_k)^{\mathrm{t}}$ by $\widehat{\boldsymbol{a}}_k = (\widehat{a}_1(k), \ldots, \widehat{a}_k(k))^{\mathrm{t}}$. The purpose of fitting such an autoregressive model is often to predict the one-step-ahead value Y_{n+1} of a new, independent realisation of the time series, denoted $\{Y_i\}$. The one-step-ahead prediction is built from the model (4.9):

$$\widehat{Y}_{n+1} = \widehat{a}_1(k) Y_n + \widehat{a}_2(k) Y_{n-1} + \cdots + \widehat{a}_k(k) Y_{n-k+1}.$$

Conditional on the original time series $X_1, \ldots, X_n$, the mean squared prediction error of $\widehat{Y}_{n+1}$ equals

$$\begin{aligned} &\mathrm{E}[(Y_{n+1} - \widehat{Y}_{n+1})^2 | X_1, \ldots, X_n] \\ &\quad = \mathrm{E}[(Y_{n+1} - a_1 Y_{n+j-1} - \cdots - a_k Y_{n+1-k} \\ &\qquad -\{(\widehat{a}_1(k) - a_1) Y_{n+1-1} - \cdots - (\widehat{a}_k(k) - a_k) Y_{n+1-k}\})^2 | X_1, \ldots, X_n] \\ &\quad = \sigma^2 + (\widehat{\boldsymbol{a}}_k - \boldsymbol{a})^{\mathrm{t}} \Gamma_k (\widehat{\boldsymbol{a}}_k - \boldsymbol{a}), \end{aligned} \tag{4.10}$$

where the (i, j)th entry of Γ_k is defined as $\mathrm{E}(Y_i Y_j)$, for $i, j = 1, \ldots, k$. For the equality in the last step, we have used the independence of $\{Y_i\}$ and $X_1, \ldots, X_n$.

Before continuing with the discussion on efficiency, we explain how (4.10) leads to Akaike's final prediction error (Akaike, 1969)

$$\mathrm{FPE}(k) = \widehat{\sigma}_k^2 \frac{n + k}{n - k}, \tag{4.11}$$

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

where $\widehat{\sigma}_k^2$ is the maximum likelihood estimator of σ^2 in the truncated model of order k, that is, $\widehat{\sigma}_k^2 = \mathrm{SSE}_k/n$. FPE is used as an order selection criterion to determine the best truncation value k, by selecting the k that minimises $\mathrm{FPE}(k)$. This criterion is directed towards selecting a model which performs well for prediction. For a modified version, see Akaike (1970).

The expression in (4.10) is a random variable. To compute its expectation with respect to the distribution of $X_1, \ldots, X_n$, we use that for maximum likelihood estimators $\widehat{\boldsymbol{a}}_k$ (see Brockwell and Davis, 1991, Section 8.8),

$$\sqrt{n}(\widehat{\boldsymbol{a}}_k - \boldsymbol{a}_k) \xrightarrow{d} \mathrm{N}(0, \sigma^2 \Gamma_k^{-1}).$$

This implies that for large n we may use the approximation

$$\sigma^2 + \mathrm{E}\{(\widehat{\boldsymbol{a}}_k - \boldsymbol{a})^{\mathrm{t}} \Gamma_k (\widehat{\boldsymbol{a}}_k - \boldsymbol{a})\} \approx \sigma^2(1 + k/n).$$

Replacing the true unknown σ^2 by the unbiased estimator $n\widehat{\sigma}_k^2/(n-k)$, leads to the $\mathrm{FPE}(k)$ expression (4.11).

To prove efficiency, some assumptions on the time series are required. We refer to Lee and Karagrigoriou (2001), who provide a set of minimal assumptions in order for efficiency to hold. The main result is the following:

Theorem 4.4 *Under the assumptions as in Lee and Karagrigoriou (2001):*

(a) The criteria $\mathrm{AIC}(k) = -n \log(\widehat{\sigma}_k^2) - 2(k+1)$*, the corrected version* $\mathrm{AIC}_c(k) = -n \log(\widehat{\sigma}_k^2) - n(n+k)/(n-k-2)$*,* $\mathrm{FPE}(k)$*,* $S_n(k) = \mathrm{FPE}(k)/n$*, and* $\mathrm{FPE}_\alpha(k) = \widehat{\sigma}_k^2 (1 + \alpha k/n)$ *with* $\alpha \neq 2$*, are all asymptotically efficient.*

(b) The criteria $\mathrm{BIC}(k) = \log \widehat{\sigma}_k^2 + k \log n$ *and the Hannan–Quinn criterion* $\mathrm{HQ}(k) = n \log \widehat{\sigma}_k^2 + 2ck \log \log n$ *with* $c > 1$ *are not asymptotically efficient.*

The assumption of predicting the next step of a new and independent series $\{Y_t\}$ is not quite realistic, since in practice the same time series is used for both selecting the order and for prediction. Ing and Wei (2005) develop an efficiency result for AIC-like criteria for what they term same-realisation predictions where the original time series is also used to predict future values. Their main conclusion is that the efficiency result still holds for AIC for the one-series case.

4.7 Efficient selection of regression variables

The onset to studying efficiency in regression variable selection is given by Shibata (1981, 1982) for the situation of a true model containing an infinite or growing number of regression variables. For a normal regression model, we follow the approach of Hurvich and Tsai (1995b) to obtain the efficiency of a number of model selection criteria.

We make the assumption that for all considered index sets S, the corresponding design matrix X_S (containing in its columns those variables x_j with j in S) has full rank $|S|$,

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

where $|S|$ denotes the number of variables in the set S, and that $\max_S |S| = o(n^a)$ for a constant $a \in (0, 1]$. There is another technical requirement that asks for the existence of a constant $b \in [0, 0.5)$ such that for any $c > 0$,

$$\sum_S \exp\{-cn^{-2b}L_n(S)\} \to 0, \quad \text{for } n \to \infty.$$

This implies in particular that for all S, $n^{-2b}L_n(S) \to \infty$. This condition is fulfilled when the number of regression variables in the true model is infinite, or for models where the number of variables grows with the sample size. One such example is a sequence of nested models,

$$S_1 = \{1\} \subset S_2 = \{1, 2\} \subset \cdots \subset S_q = \{1, \ldots, q\} \subset \cdots,$$

where $|S_q| = q$. Hurvich and Tsai (1995a) (see also Shibata, 1981) provide some ways to check the technical assumption.

The next theorem states the asymptotic efficiency of the specific model selection criterion

$$\text{IC}(S) = (n + 2|S|)\widehat{\sigma}_S^2, \tag{4.12}$$

where the variance in model S is estimated via $n^{-1}(Y - X_S\widehat{\beta}_S)^{\text{t}}(Y - X_S\widehat{\beta}_S) = \text{SSE}_S/n$. For the proof we refer to Hurvich and Tsai (1995a).

Theorem 4.5 *Let $\widehat{S}_{\text{IC}}$ denote the set selected by minimising criterion (4.12). If the assumptions hold, then, with S^* defined as the minimiser of L_n in (4.7), and $c = \min\{(1 - a)/2, b\}$,*

$$\frac{L_n(\widehat{S}_{\text{IC}})}{L_n(S^*)} - 1 = o_P(n^{-c}).$$

The criterion in (4.12) is for normal regression models closely related to AIC. It is hence no surprise that AIC has a similar behaviour, as the next corollary shows. The same can be said of Mallows's (1973) C_p, see (4.5).

Corollary 4.1 *(a) The criteria* AIC, *final prediction error* FPE, *see (4.11), their modifications* $\text{FPE}(S)/n$, $\text{AIC}_c(S) = \text{AIC}(S) + 2(|S| + 1)(|S| + 2)/(n - |S| + 2)$ *and Mallows's C_p are all asymptotically efficient under the conditions of Theorem 4.5.*

(b) The criteria BIC *and the Hannan–Quinn criterion* HQ *are not asymptotically efficient.*

This corollary follows using the method of proof of Shibata (1980, theorem 4.2). Nishii (1984) has shown that FPE and Mallows's C_p are asymptotically equivalent.

4.8 Rates of convergence*

The result in Theorem 4.5, as obtained by Hurvich and Tsai (1995b), is stronger than just showing efficiency. Indeed, it gives the rate by which the ratio of the function L_n

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

evaluated at the set chosen by the information criterion, and at the optimal set, converges to one; this rate is stated as $o_P(n^{-c})$, where $c = \min\{(1-a)/2, b\}$. For $a = 1$ and $b = 0$, the theorem reduces to the original result of Shibata (1981). In this case we have that $c = 0$ and that the ratio is of the order $o_P(1)$, or in other words, there is convergence to zero in probability. When $c = 0$ we do not know how fast the convergence to zero is. The constants a and b determine how fast the rate of convergence will be. The a value is related to the size of the largest model. The value $0 < a \leq 1$ is such that $\max_S |S| = o(n^a)$. The rate of convergence increases when a decreases from 1 to $1 - 2b$. However, a smaller value of a implies a smaller dimension of the largest model considered. This means that it cannot always be advised to set a to its smallest value $1 - 2b$, since then we are possibly too much restricting the number of parameters in the largest model. The value $0 \leq b < 1/2$ gives us information on the speed by which L_n diverges. From the assumption we learn that b is such that $n^{-2b} L_n(S) \to \infty$.

By changing the values for a and b to have $a \to 0$ and $b \to 1/2$, one obtains rates close to the parametric rate of convergence $o_P(n^{-1/2})$.

For a study on convergence rates of the generalised information criterion for linear models, we refer to Shao (1998). Zhang (1993) obtained rates of convergence for AIC and BIC for normal linear regression.

4.9 Taking the best of both worlds?*

Both AIC and the BIC have good properties, AIC is efficient and the BIC is consistent. A natural question is whether they can be combined. Bozdogan (1987) studies the bias introduced by the maximum likelihood estimators of the parameters and proposes two adjustments to the penalty of AIC. This leads to his 'corrected AIC' of which a first version is defined as $\text{CAIC} = 2\,\ell(\widehat{\theta}) - \text{length}(\theta)(\log n + 1)$. In the notation of Chapter 3, $\text{CAIC} = \text{BIC} - \text{length}(\theta)$. A second version uses the information matrix $J_n(\widehat{\theta})$ and is defined as $\text{CAICF} = 2\,\ell(\widehat{\theta}) - \text{length}(\theta)(\log n + 2) - \log|J_n(\widehat{\theta})|$. Bozdogan (1987) shows that for both corrected versions the probability of overfitting goes to zero asymptotically, similarly as for the BIC, while keeping part of the penalty as in AIC, namely a constant times the dimension of θ.

A deeper question is whether the consistency of the BIC can be combined with the efficiency of AIC. Yang (2005) investigates such questions and comes to a negative answer. More precisely, he investigates a combination of consistency and minimax rate optimality properties in models of the form $Y_i = f(x_i) + \varepsilon_i$. A criterion is minimax rate optimal over a certain class of functions $\mathcal{C}$ when the worst case situation for the risk, that is $\sup_{f \in \mathcal{C}} n^{-1} \sum_{i=1}^n \text{E}[\{f(x_i) - f_{\widehat{S}}(x_i, \widehat{\theta}_{\widehat{S}})\}^2]$, converges at the same rate as the best possible worst case risk

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{C}} n^{-1} \sum_{i=1}^n \text{E}\big[\{\widehat{f}(x_i) - f(x_i)\}^2\big].$$

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

In the above, the selected model is denoted $f_{\widehat{S}}$, while $\widehat{f}$ is any data-based estimator of f. AIC can be shown to be minimax rate optimal, while the BIC does not have this property. Changing the penalty constant 2 in AIC to some other value takes away this favourable situation. Hence, just changing the penalty constant cannot lead to a criterion that has both properties. In theorem 1 of Yang (2005), he proves the much stronger result that any consistent model selection method cannot be minimax rate optimal. Model averaging (see Chapter 7) is also of no help here; his theorem 2 shows that any model averaging method is not minimax rate optimal if the weights are consistent in that they converge to one in probability for the true model, and to zero for the other models. Theorem 3 of Yang (2005) tells a similar story for Bayesian model averaging.

4.10 Notes on the literature

Results about theoretical properties of model selection methods are scattered in the literature, and often proofs are provided for specific situations only. Exceptions are Nishii (1984) who considered nested, though possibly misspecified, likelihood models for i.i.d. data, and Sin and White (1996). Consistency (weak and strong) of the BIC for data from an exponential family is obtained by Haughton (1988, 1989). Shibata (1984) studied the approximate efficiency for a small number of regression variables, while Li (1987) obtains the efficiency for Mallows's C_p, cross-validation and generalised cross-validation. Shao and Tu (1995, Section 7.4.1) show the inconsistency of leave-one-out cross-validation. Breiman and Freedman (1983) construct a different efficient criterion based on the expected prediction errors. A strongly consistent criterion based on the Fisher information matrix is constructed by Wei (1992). Strongly consistent criteria for regression are proposed by Rao and Wu (1989) and extended to include a data-determined penalty by Bai *et al.* (1999). An overview of the asymptotic behaviour of model selection methods for linear models is in Shao (1997). Zhao *et al.* (2001) construct the efficient determination criterion EDC, which allows the choice of the penalty term d_n, to be made over a wide range. In general their d_n can be taken as a sequence of positive numbers depending on n or as a sequence of positive random variables. Their main application is the determination of the order of a Markov chain with finite state space. Shen and Ye (2002) develop a data-adaptive penalty based on generalised degrees of freedom. Guyon and Yao (1999) study the probabilities of underfitting and overfitting for several model selection criteria, both in regression and autoregressive models. For nonstationary autoregressive time series models, weak consistency of order selection methods is obtained by Paulsen (1984) and Tsay (1984). For strong consistency results we refer to Pötscher (1989).

The consistency property has also been called 'the oracle property', see e.g. Fan and Li (2002). The effect of such consistency on further inference aspects has been studied by many authors. One somewhat disturbing side-effect is that the max-risk of post-selection estimators divided by the max-risk of ordinary estimators may

diverge to infinity. Versions of this phenomenon have been recognised by Foster and George (1994) for the BIC in multiple regression, in Yang (2005) as mentioned above, in Leeb and Pötscher (2005, 2006, 2008) for subset selection in linear regression models, and in Hjort and Claeskens (2006) for proportional hazards models. Importantly, various other nonconsistent selection methods, like AIC and the FIC (Chapter 6), are immune to this unbounded max-risk ratio problem, as discussed in Hjort and Claeskens (2006), for example. The Leeb and Pötscher articles are also concerned with other aspects of post-model-selection inference, like the impossibility of estimating certain conditional distributions consistently. These aspects are related to the fact that it is not possible to estimate the δ parameter consistently in the local $\delta/\sqrt{n}$ framework of Chapter 5.

Exercises

4.1 *Frequency of selection:* Perform a small simulation study to investigate the frequency by which models are chosen by AIC, the BIC and the Hannan–Quinn criterion. Generate (independently for $i = 1, \ldots, n$)

$$x_{i,1} \sim \text{Uniform}(0, 1),\ x_{i,2} \sim N(5, 1),\ Y_i \sim N(2 + 3x_{i,1}, (1.5)^2).$$

Consider four normal regression models to fit:
M_1: $Y = \beta_0 + \varepsilon$
M_2: $Y = \beta_0 + \beta_1 x_1 + \varepsilon$
M_3: $Y = \beta_0 + \beta_2 x_2 + \varepsilon$
M_4: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
For sample sizes $n = 50$, 100, 200, 500 and 1500, and 1000 simulation runs, construct a table which for each sample size shows the number of times (out of 1000 simulation runs) that each model has been chosen. Do this for each of AIC, the BIC and Hannan–Quinn. Discuss.

4.2 *Hofstedt's highway data:* Consider the data of Exercise 2.10. Use robust C_p to select variables to be used in the linear regression model. Construct a plot of the values of C_p versus the variable V_p. Compare the results to those obtained by using the original version of C_p.

4.3 *Calculating the risk for the choice between two normal models:* Let $Y_1, \ldots, Y_n$ be i.i.d. from the N(μ, 1) density, as in Section 4.2.
(a) For the estimator (4.2), show that

$$\sqrt{n}(\widehat{\mu} - \mu) = Z(1 - A_n) - \sqrt{n}\mu A_n,$$

where $A_n = I\{|\sqrt{n}\mu + Z| \le d_n^{1/2}\}$, and Z is a standard normal variable.
(b) Show that the risk function $r_n(\mu)$, defined as n times mean squared error,

$$r_n(\mu) = n\,\mathrm{E}_\mu\{(\widehat{\mu} - \mu)^2\} = \mathrm{E}\big[\big\{(\sqrt{n}\mu + Z)I\big\{|\sqrt{n}\mu + Z| \ge d_n^{1/2}\big\} - \sqrt{n}\mu\big\}^2\big],$$

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press

can be written as

$$r_n(\mu) = 1 - \int_{\text{low}_n}^{\text{up}_n} z^2 \phi(z)\, \mathrm{d}z + n\mu^2 \{\Phi(\text{up}_n) - \Phi(\text{low}_n)\},$$

where

$$\text{low}_n = -d_n^{1/2} - \sqrt{n}\mu \quad \text{and} \quad \text{up}_n = d_n^{1/2} - \sqrt{n}\mu.$$

Plot the resulting risk functions corresponding to the AIC and BIC methods for different sample sizes n.

https://doi.org/10.1017/CBO9780511790485.005 Published online by Cambridge University Press