IMA Journal of Mathematical Control & Information (1986) **3**, 211–222

A Predictive Least-Squares Principle

JORMA RISSANEN
IBM Res., K54/802, San Jose, Ca. 95120-6099

[Received 2 October 1985 and in revised form 16 July 1986]

A new principle of least-squares estimation is described, which extends the old in allowing the estimation of the number of the parameters along with their values. Just as the old principle, the new one too uses only a sum of squares, which now, however, represent prediction errors rather than fitting errors. In a typical regression problem with independent normal data, the estimates of the number of the parameters, i.e. the number of the regressor functions, are shown to be consistent.

1. Introduction

THE so-called selection-of-variables problem in regression is one of the fundamental problems in statistics, and a large number of criteria such as C_p (Mallows, 1974), MSEP (Allen, 1971), IMSE (Helms, 1974), AIC (Akaike, 1974), Cross-Validation (Stone, 1974, 1977; Geisser & Eddy, 1979; Picard & Cook, 1984), to mention only the better ones, has been proposed for its solution. But unlike the simpler regression problem with a fixed known number of variables, where the Gauss–Markov theorem settles the issue of preferred solutions in favour of the least-squares principle, there is no rational basis to compare these criteria—a situation which we take as symptomatic of the state of today's statistics. That none of the proposed criteria qualifies as a serious candidate for an optimal solution in any meaningful sense is evidenced by the fact that all fail a basic test of performance: The estimates of the number of parameters or, equivalently, the number of regression variables, in the basic linear gaussian regression problem are not consistent.

In this paper we study as an application of the notion of 'predictive stochastic complexity' (Rissanen, 1986a) a new estimation principle for the selection-of-variables problem. Just as the least-squares principle of Gauss, the new one also uses only squared deviations, and we call it the predictive least-squares principle (PLS). Unlike the criteria cited above, which cannot be given any natural meaning, our minimized criterion admits two fundamental interpretations. Firstly, it represents the least total of accumulated prediction errors, when the observed data are predicted one by one such that each prediction is done with knowledge of the so-far processed observations, only. Secondly, the same sum also gives, to within a constant, the greatest lower bound, i.e. the stochastic complexity, of the number of binary digits with which the predicted sequence can be encoded from the observed values of the regressor variables with use of an appropriate code. Both bounds, then, provide the rational basis for comparing different solutions to the selection-of-variables problem, as studied in Rissanen

(1986a, b). In this paper, as the main result, we prove in the basic Gaussian case that the predictive least-squares estimates of the number of the parameters are consistent.

The predictive least-squares principle is related to the 'prequential' principle in Dawid (1984), which is applicable to naturally ordered sequences, and also to the 'forward validation' idea in Hjorth (1982), where, however, it was used in a traditional way to reduce bias. The wider coding-theoretic view of modelling is also discussed in Rissanen (1978, 1983).

2. Accumulated prediction errors

We consider the basic regression problem, where we try to 'explain' a sequence of n observations $y_1, \ldots, y_n$ of a variable y in terms of K other variables $x_1, \ldots, x_K$, each observed at the corresponding n points. Intuitively speaking, we want a measure of the degree of 'explanation' such that it allows us to estimate or predict well the future values y_t $(t > n)$ from the corresponding observations x_{it} $(i = 1, \ldots, K)$ of the regressor variables. In general, the best predictor will be a function of a k-element subcollection of the regressor variables,

$$\hat{y}_t = f_\theta(x_{i(1),t}, \ldots, x_{i(k),t}), \tag{2.1}$$

where θ is a vector parameter with k components. Any minimization procedure calls for a search through all the subcollections of the regressor variables in order to find the best, which is a task of exponential complexity. However, there is no alternative, because clearly we cannot find the best collection without having tried them all. Whether we consider the search through all subcollections or only over the k first for $k = 0, 1, \ldots$, is irrelevant as far as the principle that we shall describe is concerned, and for the sake of simplicity we take $x_{i(j)} = x_j$, which means that we only search for the optimal value for k.

At this point it is generally assumed in the literature that there is a parent distribution connecting the y variable with the regressor variables such that we can write

$$y_t = g(x_{1,t}, \ldots x_{K,t}) + e_t, \tag{2.2}$$

where e_t is an uncorrelated sequence with zero mean. Moreover, the function g is in the same family as (2.1), so that it is defined by a 'true' parameter $\bar{\theta}$. Then the ideal predictor should minimize $\mathrm{E}_{\bar{\theta}}(y_t - \hat{y}_t)^2$ for the future values of t and for all 'true' parameters (Allen, 1971). Although it is realized that no such 'true' parameter really exists, it is invented to provide a convenient target for estimation. Hence, if we replace the expectation by an estimate and construct the predictor apropriately in terms of the estimated parameters, we get something which may be taken as an approximation of the ideal predictor. Depending on how this is done, the various criteria, some of which were mentioned above, result. The main problem with this thinking is that it involves a number of subjective choices, and that it is quite impossible to interpret the final result in a meaningful data-dependent manner, which would allow a rational comparison of

estimators. All one can say for sure is that the end result is some kind of approximation of something that does not exist.

We take a fundamentally different approach which, although quite simple, has its subtle points (Rissanen, 1986a). We redescribe it here in a self-containing manner for the quadratic error measure which we are interested in. We begin with the simpler case where the observations are regarded as a sequence, ordered by the index t, which we take to be time instance, so that we can speak of the 'past' data points. Without requiring the existence of any 'true' distributions, we certainly may consider a predictor

$$\hat{y}_t = f_{\hat{\theta}(t-1)}(x_{1,t}, \ldots, x_{k,t}),$$

where for each number of components k the parameter $\hat{\theta}(t-1)$ is restricted to be a suitable function of the past data $(y_j, x_{1,j}, \ldots, x_{k,j})$ for $j = 1, \ldots, t-1$. This makes the predictions 'honest' in that they will have to be made from the available (past) information. How should we pick the estimator $\hat{\theta}(t-1)$? We would, of course, like this estimator to be optimal in some desired direction, the most obvious one being so that the prediction error $(y_t - \hat{y}_t)^2$ is minimized. But this would make the estimate a function of y_y, in violation of the 'honesty' requirement. We see then that we have a problem, in fact, so much so that we must abandon the attempt to find an optimized estimator, and instead settle for less. We resort to intuition and just pick the estimator $\hat{\theta}(t-1)$ for each k such that it minimizes the *past* squared deviations

$$R(t-1, k) = \min_{\theta} \sum_{j=1}^{t-1} [y_j - f_\theta(x_{1j}, \ldots, x_{kj})]^2. \tag{2.3}$$

This seems a reasonable choice. Indeed, any other choice would raise the question: why not pick the single value of the parameter which would have worked best in the past, had we the foresight to use it? We then accumulate all the resulting prediction errors

$$I(y \mid x, k) = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2, \tag{2.4}$$

which may be minimized over k with the result $\hat{k}(n)$. If there are several minimizing values, we pick the smallest. With (2.3) this defines the final estimate $\hat{\theta}(n)$. We call these the PLS estimates.

The expression (2.4) requires the estimates $\hat{\theta}(t)$ even for $t < k$, and in particular for $t = 0$ when no data are available. In lack of prior knowledge, we take $\hat{y}_1 = 0$ for all values for k. Further, whatever value for k is chosen, we for each $t < k$ compute from (2.3) only as many components in $\hat{\theta}(t-1)$ as can be solved uniquely. In other words, as t grows, we gradually increase the number of parameters one by one until the set value, with which (2.4) is to be evaluated, is reached.

We now describe the case where the unordered list of data, rather than some particular sequence, is to be predicted. This is important in cases where the data are modelled as being independent so that no information is expected from the

order. In principle, we should consider the minimum of $I(y \mid x, \hat{k}(n))$ over all permuted sequences, but this, of course, is too formidable a task even for moderate-sized n. Instead, we construct a symmetric function of the data more simply as follows. In the absence of prior knowledge, suppose that we predict the very first observation as zero. Which observation should this be? We pick the one that results in the smallest error; i.e., one that has the smallest absolute value. Call the index of this $i(1)$. Repeating this, we get a particular sequence with the formula

$$I(y \mid x, k) = \sum_{t=0}^{n-1} \min_{j \notin \{i(1), \ldots, i(t)\}} [y_i - \hat{f}_{\hat{\theta}(t)}(x_{1,j}, \ldots, x_{kj})]^2, \tag{2.5}$$

where $\hat{\theta}(t)$ is a function of the so-far processed data $y_{i(s)}, x_{1,i(s)}, \ldots, x_{k,i(s)}$, for $s = 1, \ldots, t$. The minimizing index j defines $i(t+1)$.

We conclude this section with a discussion of the central ideas involved. First, it is our thesis that this predictive least-squares principle is hard to beat. For one thing, if there is any machinery behind the data that restricts the future observation in a manner it did in the past and which can be captured by the selected class of parametric functions, then we will find that machinery. If, again, no such machinery exists, then our predictions will be bad, but so will all other predictions, too, that use the same class of parametric functions. Further, the criterion we seek to minimize expresses exactly the right thing, namely, the prediction errors on the observations, rather than the mean prediction error, the mean taken relative to some hypothetical 'true' distribution. Finally, the principle involves few arbitrary choices that need to be made by 'sound judgment'. One is the selection of the parametric class; however, this is inevitable. Another is the agreement to use quadratic measures, which could be dispensed with. The only really arbitrary selection is to use the least-squares estimates (2.3), which may be then taken as a good and a rather natural guess. Notice that the fact that these estimates minimize the quadratic residuals in the past does not imply that they will minimize the squared prediction errors in (2.5), which is what really counts, but here we are facing the essence of inductive inference, and no entirely satisfactory optimized procedure for doing it can ever be formulated.

3. Linear regression

We are particularly interested in fitting linear functions,

$$y_t = \theta_1 x_{1,t} + \cdots + \theta_k x_{k,t} + e_t \quad (t = 1, \ldots, n), \tag{3.1}$$

where e_t represents the amount by which the linear model fails to explain y_t. Introduce the notation

$$\boldsymbol{X}^{\mathsf{T}}(t, k) = [x_{ij}: i = 1, \ldots, k; j = 1, \ldots, t], \qquad \boldsymbol{C}(t, k) = \boldsymbol{X}^{\mathsf{T}}(t, k)\boldsymbol{X}(t, k), \tag{3.2}$$

where T denotes transposition. We also denote the bottom k-element row of the $t \times k$ matrix $\boldsymbol{X}(t, k)$ by $\boldsymbol{x}^{\mathsf{T}}(t, k) = [x_{1,t}, \ldots, x_{k,t}]$ and write $\boldsymbol{y}^{\mathsf{T}}(t) = [y_1, \ldots, y_t]$. Then the ordinary least-squares solution $\hat{\boldsymbol{\theta}}(t, k)$ with k components, when fitted

to the data $y_1, \ldots, y_t$, is given by

$$\hat{\boldsymbol{\theta}}(t, k) = \boldsymbol{C}^{-1}(t, k)\boldsymbol{X}^{\mathrm{T}}(t, k)\boldsymbol{y}(t) \quad (t \geq t_k), \tag{3.3}$$

where t_k denotes the smallest integer such that the matrix $\boldsymbol{C}$ has inverse. Let

$$e_t = y_t - \hat{\boldsymbol{\theta}}^{\mathrm{T}}(t-1, k)\boldsymbol{x}(t, k) \tag{3.4}$$

denote the error when y_t is predicted as a linear combination of the k first regressor functions by use of the least-squares solution, defined by the past data up to time $t-1$. By summing up the squares of these errors we get (2.4) in the form

$$I(y \mid x, k) = \frac{1}{n}\sum_{t=1}^{n} e_t^2. \tag{3.5}$$

The minimized expression $I(y \mid x, \hat{k}(n))$ differs from the sum of the minimized squared residuals $R(n, k) = 1/n[\boldsymbol{y}^{\mathrm{T}}(n)\boldsymbol{y}(n) - \boldsymbol{y}^{\mathrm{T}}(n)\boldsymbol{X}(n, k)\hat{\boldsymbol{\theta}}(n, k)]$ in that it depends on the sequence of the least-squares estimates $\hat{\boldsymbol{\theta}}(1, 1), \ldots, \hat{\boldsymbol{\theta}}(n-1, k)$ rather than on just the last $\hat{\boldsymbol{\theta}}(n, k)$. This gives another intuitive reason why the criterion (3.5) and not $R(n, k)$ can be minimized over k with a meaningful result. Indeed, while $R(n, k)$ steadily decreases as k increases, the early parameter estimates in the given sequence tend to be poor and more so as k grows, which implies that $I(y \mid x, k)$ need not, and in general will not, be minimized for the largest possible value for k, namely, n. The case with ARMA processes is handled as a special case with the regressor variables as suitable shifts of the main variables y_t. Such modelling is studied in Rissanen (1986b).

We illustrate the PLS estimates with an example from Picard & Cook (1984). From 20 flocks of Canada geese the numbers x_i $(i = 1, \ldots, 20)$ of adult birds were estimated as 10, 10, 12, 20, 40, 40, 30, 30, 20, 20, 18, 35, 35, 35, 30, 50, 30, 30, 45, and 30, respectively. The same flocks were also photographed, from which the true numbers of adult birds y_i $(i = 1, \ldots, 20)$ were counted. Written in the same order as the corresponding estimates, they are: 9, 11, 14, 26, 57, 56, 38, 38, 22, 22, 18, 43, 42, 42, 34, 62, 30, 30, 48, and 25. We wish to fit a polynomial predictor to predict y_t from x_t, perhaps in order to avoid taking expensive photographs, thus:

$$y_t = \theta_1 + \theta_2 x_t + \cdots + \theta_k x_t^{k-1},$$

where the number of parameters is to be estimated, too. In this case the elements of the matrix $\boldsymbol{X}^{\mathrm{T}}(t, k)$ are given as $x_{ij} = x_j^{i-1}$ $(i = 1, \ldots, k; j = 1, \ldots, n)$.

Solving the unordered minimization problem (2.5), we get to the indicated precision $I(y \mid x, 1) = 4930$, $I(y \mid x, 2) = 548$, $I(y \mid x, 3) = 597$, $I(y \mid x, 4) = 1182$, from which we conclude that the minimizing polynomial is linear. Its coefficients are given by $\theta_1 = -3.8$, $\theta_2 = 1.3$, and the resulting line fits the data well. In fact, there is no doubt at all that a human observer, using his judgement, would have picked the same line as the best fit. The given two sequences correspond to the optimum order, defined by (2.5).

4. Consistency

The PLS estimates in the preceding section were derived from what appears to us to be a most reasonable inductive principle, and the associated estimates may be expected to be good, if the future behaves like the past. Hence, if we wish to prove good properties of the estimators, we must make sure that the future indeed behaves like the past and on the whole consider a situation which we can analyze. We study the basic regression problem, about which we make the following assumptions. First, we let the data be generated by independent, identically $N(0, \sigma)$-distributed variables, ε_t, as follows

$$y_t = \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}(t, m) + \varepsilon_t. \tag{4.1}$$

Whatever the origin of the regressor variables, we regard them as a deterministic sequence, so that the expectation operations to be considered are with respect to the random variables ε_t. The mean of y_t is $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}(t, m)$ for each t, where both $\boldsymbol{\theta}$ and the number of its components m are unknown and to be estimated. The variance σ^2 is not estimated here, and since it plays no role in the analysis, we set $\sigma = 1$. Finally, we must assume an appropriate behaviour of the regressor variables, whose nature varies greatly from case to case. We assume that for all N and k

$$\frac{1}{M}\sum_{t=N+1}^{N+M} \boldsymbol{x}(t, k)\boldsymbol{x}^{\mathrm{T}}(t, k) \to \boldsymbol{C}(k) \quad \text{as } M \to \infty, \qquad \boldsymbol{C}(k) > 0. \tag{4.2}$$

This is seen to be a stationarity assumption of sorts, in particular, when the values of the regressor variables were themselves a sample from a random process. The condition can be seen to hold for example in the polynomial case, if $\boldsymbol{x}_t$ remains uniformly bounded.

We state the main result.

THEOREM 1 *Under the stated assumptions and with finitely many regressor variables the* PLS *estimate* $\hat{k}(n)$ *is a consistent estimate of* m, *the number of components in the parameter vector* $\boldsymbol{\theta}$ *of the data generating process* (4.1); *i.e.* $\hat{k}(n) \to m$ *in probability.*

The proof is given in Appendix B, which also uses the results in Appendix A.

Appendix A

We begin with a few formulas. Partition the parameter vector as

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}(k) \\ \boldsymbol{\theta}(k, m) \end{bmatrix},$$

where the first part denotes the first k components and the second the remaining $m - k$. Corresponding partitioning is used for the various other vectors and matrices; in particular, $\boldsymbol{Z}(t, k, m)$ denotes the matrix defined by the last $m - k$ columns of $\boldsymbol{X}(t, m)$, and

$$\boldsymbol{z}^{\mathrm{T}}(t, k, m) = [x_{k+1,t}, \ldots, x_{m,t}].$$

This is done in order to avoid writing essentially the same equations twice, namely, for the case $k < m$ and for the remaining case. We then obtain, from (3.3)–(3.4) and (4.1) for $t \geq t_k$, the result

$$e_t - \lambda_t(k) = \varepsilon_t - \boldsymbol{x}^{\mathsf{T}}(t, k)\boldsymbol{C}^{-1}(t-1, k)\sum_{i<t}\boldsymbol{x}(i, k)\varepsilon_i = \varepsilon_t - \sum_{i<t}\psi_{t,i}\varepsilon_i, \qquad \text{(A.1)}$$

where the mean $\mathsf{E}e_t = \lambda_t(k)$ is given by

$$\lambda_t(k) = [\boldsymbol{z}^{\mathsf{T}}(t, k, m) - \boldsymbol{x}^{\mathsf{T}}(t, k)\boldsymbol{C}^{-1}(t-1, k)\boldsymbol{X}^{\mathsf{T}}(t-1, k)\boldsymbol{Z}(t-1, k, m)]\boldsymbol{\theta}(k, m). \qquad \text{(A.2)}$$

The mean $\lambda_t(k)$ is seen to be zero when $k \geq m$. From (A.1) and the definition of $\boldsymbol{C}(t, k)$ we deduce with straightforward calculations that the zero-mean gaussian sequence $e_t - \lambda_t(k)$ is independent for each k with the variance

$$\mathsf{E}[e_t - \lambda_t(k)]^2 = 1 + \mu_t(k), \qquad \text{(A.3)}$$

where

$$\mu_t(k) = \boldsymbol{x}^{\mathsf{T}}(t, k)\boldsymbol{C}^{-1}(t-1, k)\boldsymbol{x}(t, k). \qquad \text{(A.4)}$$

LEMMA 1 *For every positive ε, the sum*

$$\alpha_n(k) = \frac{1}{n}\sum_{t=1}^{n}\mu_t(k)$$

satisfies the inequalities

(i) $$(k-\varepsilon)\frac{\ln n}{n} \leq \alpha_n(k) \leq (k+\varepsilon)\frac{\ln n}{n}$$

for all n large enough. Further, there is a positive constant c such that

(ii) $$\frac{1}{n}\sum_{t=1}^{n}\mu_t^2(k) < c\frac{\ln n}{n},$$

for all n large enough. Finally, for $k < m$ we have

(iii) $$\beta_n(k) = \frac{1}{n}\sum_{t=1}^{n}\lambda_t^2(k) \to c_k, \qquad c_k > 0,$$

for all n large enough; in the remaining case $k \geq m$, we have $\lambda_t(k) = 0$ for $t \geq t_m$, and $\beta_n(k) \leq c'/n$ for all n large enough and for some positive constant c'.

Proof. The proof is straightforward, and we just indicate the major steps. We keep k fixed and suppress it in the following expressions except in $\boldsymbol{x}(t, k)$, where its absence might cause confusion. Divide the interval $1, \ldots, n$ into intervals of length M, which eventually is selected large, and let s denote the smallest integer such that $n \leq sM$. With the notation

$$\boldsymbol{G}_{r,M} := \frac{1}{M}\sum_{t=rM+1}^{(r+1)M}\boldsymbol{x}(t, k)\boldsymbol{x}^{\mathsf{T}}(t, k),$$

we deduce from the assumption (4.2) that

$$C^{-1}(t) =: \frac{1}{t} C^{-1} + \frac{1}{t} \Gamma_t, \qquad G_{r,M} =: C + \Delta_{r,M}, \tag{A.5}$$

where $\Gamma_t \to 0$ as $t \to \infty$, and $\Delta_{r,M} \to 0$ as $M \to \infty$.

We have further

$$\begin{aligned} R_n &= \operatorname{tr} \sum_{t=1}^{n} C^{-1}(t) x(t+1, k) x^{\mathrm{T}}(t+1, k) \\ &= \operatorname{tr} C^{-1} \sum_{t=1}^{n} (I + C\Gamma_t) \frac{1}{t} x(t+1, k) x^{\mathrm{T}}(t+1, k) \\ &\leqslant \operatorname{tr} C^{-1} \sum_{t=1}^{M} (I + C\Gamma_t) x(t+1, k) x^{\mathrm{T}}(t+1, k) + \\ &\qquad \operatorname{tr} C^{-1} \sum_{r=1}^{s} \frac{1}{rM} \sum_{t=rM+1}^{(r+1)M} (I + C\Gamma_t) x(t+1, k) x^{\mathrm{T}}(t+1, k). \end{aligned}$$

We now use (A.5) and the estimate

$$\sum_{r=1}^{s} \frac{1}{rM} \leqslant \frac{1}{M} \ln n + 1 + \frac{1}{sM}$$

and separate the terms involving Γ_t and $\Delta_{r,M}$ from the rest. By picking M and n large enough, we conclude that

$$\alpha_n(k) = \frac{1}{n} \operatorname{tr} R_n$$

satisfies the second inequality in (i) of Lemma 1. If we further use the estimates

$$\sum_{t=rM+1}^{(r+1)M} \frac{1}{t} x(t+1, k) x^{\mathrm{T}}(t+1, k) \geqslant \frac{1}{(r+1)M} \sum_{t=rM+1}^{(r+1)M} x(t+1, k) x^{\mathrm{T}}(t+1, k)$$

and

$$\sum_{r=1}^{s-1} \frac{1}{rM} \geqslant \frac{1}{M} \ln n,$$

the first inequality in (i) of Lemma 1 follows.

To prove (ii) consider for $r > 1$ the inequality

$$\begin{aligned} &\operatorname{tr} \sum_{t=rM+1}^{(r+1)M} C^{-1}(t) x(t+1, k) x^{\mathrm{T}}(t+1, k) \\ &\qquad = \operatorname{tr} C^{-1} \sum_{t=rM+1}^{(r+1)M} (I + C\Gamma_t) \frac{1}{t} x(t+1, k) x^{\mathrm{T}}(t+1, k) \leqslant \operatorname{tr} \frac{1}{r} (1 + S_M) I, \end{aligned}$$

where $S_M \to 0$ as $M \to \infty$. This inequality gives us that, for M large enough, $\sum_{t=rM+1}^{(r+1)M} \mu_{t+1} < 1$ for all r greater than some number r_0. Hence

$$\sum_{t=rM+1}^{(r+1)M} \mu_{t+1}^2 < \left(\sum_{t=rM+1}^{(r+1)M} \mu_{t+1} \right)^2 < \sum_{t=rM+1}^{(r+1)M} \mu_{t+1}.$$

and so

$$\sum_{t=n_0}^{s} \mu_{t+1}^2 < \sum_{t=n_0}^{s} \mu_{t+1}.$$

This with the second inequality in (i) implies (ii).

To prove (iii) of Lemma 1 we drop the indices k and m in (A.2) and observe that the right-hand side

$$\beta_n(k) = \frac{1}{n} \operatorname{tr} \{[\boldsymbol{Z}^{\mathsf{T}}(n)\boldsymbol{Z}(n) - \boldsymbol{Z}^{\mathsf{T}}(n)\boldsymbol{X}(n)\boldsymbol{C}^{-1}(n-1)\boldsymbol{X}^{\mathsf{T}}(n-1)\boldsymbol{Z}(n-1)]\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}\}$$

converges by (4.2) to a positive number.

Appendix B

Proof of Theorem 4.1. Although the proof involves rather cumbersome details, we get remarkably simple formulas for the two crucial variances. Instead of working directly with the random variable (3.5), written now as $I_n(k)$, and its minimized value, it turns out to be far simpler to manipulate the random variable

$$\xi_n(k) = \frac{1}{n} \sum_{t=1}^{n} (e_t^2 - \varepsilon_t^2) = I_n(k) - \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t^2,$$

which clearly is seen to reach its minimum at the same value $\hat{k}(n)$ as $I_n(k)$. We calculate first the variance of the related random variable, defined for each k and n:

$$\xi_n(k) = \frac{1}{n} \sum_{t=1}^{n} \{[e_t - \lambda_t(k)]^2 - \varepsilon_t^2\}.$$

The result is then used to calculate the variance of $\xi_n(k)$, which turns out to be so small in comparison with the difference of the mean of $\xi_n(k)$ for any two adjacent values of k, that an application of Chebyshev's inequality will prove the claim.

From (A.1) and (A.3), with the notations of (A.4), we get

$$n^2 \mathsf{E}\xi_n^2(k) = \sum [T_1(t, s) + T_2(t, s) - 2T_3(t, s)].$$

The first term $T_1(t, s) = \mathsf{E}(\bar{e}_t^2 \bar{e}_s^2)$, where $\bar{e}_t = e_t - \lambda_t(k)$, is given by

$$T_1(t, s) = \begin{cases} (1 + \mu_t)(1 + \mu_s) & \text{if } t \neq s, \\ 3(1 + \mu_t)^2 & \text{if } t = s. \end{cases}$$

The second term $T_2(t, s) = \mathsf{E}(\varepsilon_t^2 \varepsilon_s^2)$ is 1 if $t \neq s$, and 3 if $t = s$. We write the third term in the notation of (A.1) as

$$T_3(t, s) = \mathsf{E}(\varepsilon_t^2 \bar{e}_s^2) = \mathsf{E}\left[\varepsilon_t^2 \left(\varepsilon_s - \sum_{l<s} \psi_{s,l} \varepsilon_l\right)^2\right].$$

For $t = s$ this term is given by

$$T_3(t, t) = 3 + \sum_{i<t} \psi_{t,i}^2 = 3 + \mu_t(k).$$

Similarly, for $t > s$, we have $T_3(t, s) = 1 + \mu_s(k)$, while for $t < s$, it is given by

$$T_3(t, s) = 1 + 3\psi_{s,t}^2 + \sum_{i \neq t} \psi_{s,i}^2 = 1 + \mu_s(k) + 2\psi_{s,t}^2.$$

Adding all, we get

$$n^2 \mathsf{E}\xi_n^2(k) = 2 \sum_{t \leq n} \mu_t^2(k) + \sum_{t,s \leq n} \mu_t(k)\mu_s(k).$$

Finally, then, for the variance $\operatorname{var} \xi_n(k)$ we get the simple formula

$$\mathsf{E}[\xi_n(k) - \mathsf{E}\xi_n(k)]^2 = \frac{2}{n^2} \sum_{t \leq n} \mu_t^2(k). \tag{B.1}$$

We also give the asymptotic upper bound

$$\mathsf{E}[\xi_n(k) - \mathsf{E}\xi_n(k)]^2 < 2c \frac{\ln n}{n^2}, \tag{B.2}$$

for all large enough n, which follows by (ii) of Lemma 1.

We use (B.1) to get the variance of $\xi_n(k)$. We have $\xi_n(k) = \bar{\xi}_n(k) - \beta_n(k) - \delta_n(k)$, where $\beta_n(k)$ is defined in Lemma 1 and

$$\delta_n(k) = \frac{2}{n} \sum_{t=1}^{n} \lambda_t(k)[e_t - \lambda_t(k)].$$

The mean of $\delta_n(k)$ is zero, and

$$\mathsf{E}\,\delta_n^2(k) = \begin{cases} \dfrac{4}{n^2} \displaystyle\sum_{t=1}^{t_m} \lambda_t^2(k)[1 + \mu_t(k)] \quad (k \geq m), & \text{(B.3)} \\ \dfrac{4}{n^2} \displaystyle\sum_{t=1}^{n} \lambda_t^2(k)[1 + \mu_t(k)] \quad (k < m). & \text{(B.4)} \end{cases}$$

The final variance is then given by a direct calculation in both cases by

$$\operatorname{var} \xi_n(k) = \operatorname{var} \bar{\xi}_n(k) + \mathsf{E}\,\delta_n^2(k) + 2\mathsf{E}\{\delta_n(k)[\xi_n(k) - \mathsf{E}\xi_n(k)]\}, \tag{B.5}$$

where the last term is seen to vanish. For $k \geq m$, in view of (B.3), which is of order $1/n^2$, an asymptotic upper bound for this variance is given by (B.2), while in the remaining case we get from (B.2), (B.4), and Lemma 1 the upper bound c/n.

We are now ready to complete the proof. Consider the event $\{\hat{k}(n) \neq m\}$, which is seen to be a subset of

$$D_n = \bigcup_{k \neq m} \{\bar{\xi}_n(k) \leq \bar{\xi}_n(m)\}.$$

Clearly,

$$\mathsf{P}\{\hat{k}(n) \neq m\} \leq \mathsf{P}(D_n) \leq \sum_{k \neq m} \mathsf{P}\{\bar{\xi}_n(k) \leq \bar{\xi}_n(m)\}.$$

Since the upper diagonal half plane $\{(x, y): x \leqslant y\}$ is a subset of $\{(x, y): x \leqslant a\} \cup \{(x, y): a \leqslant y\}$ for all numbers a, we have

$$\mathsf{P}(D_n) \leqslant \mathsf{P}\left\{\xi_n(m) \geqslant (m + \tfrac{1}{2})\frac{\ln n}{n}\right\} + \sum_{k \neq m} \mathsf{P}\left\{\xi_n(k) \leqslant (m + \tfrac{1}{2})\frac{\ln n}{n}\right\}. \quad \text{(B.6)}$$

Further,

$$\mathsf{E}\xi_n(k) = \alpha_n(k) + \beta_n(k),$$

which for $k > m$ is, by Lemma 1, greater than $(m - \frac{1}{2})(\ln n)/n$ for all large enough n. Therefore, for each $k > m$, the corresponding summand of the series in (B.6) has the upper bound

$$\mathsf{P}\left\{|\mathsf{E}\xi_n(k) - \xi_n(k)| \geqslant \left|\mathsf{E}\xi_n(k) - (m + \tfrac{1}{2})\frac{\ln n}{n}\right|\right\}.$$

Note that

$$\left|\mathsf{E}\xi_n(k) - (m + \tfrac{1}{2})\frac{\ln n}{n}\right| \geqslant \frac{(1 - \varepsilon)\ln n}{2n}$$

for $k > m$, and hence

$$\mathsf{P}\left\{|\xi_n(k) - \mathsf{E}\xi_n(k)| \geqslant \frac{(1 - \varepsilon)\ln n}{2n}\right\} \quad \text{(B.7)}$$

gives a larger upper bound for each summand of the series in (B.6) for $k > m$. In the same manner we see that (B.7) for $k = m$ gives an upper bound also for the first term in (B.6). The case $k < m$ remains. This time, because by (iii) of Lemma 1, $\mathsf{E}\xi_n(k) \geqslant \beta_n(k) \geqslant C$ for some constant C and for all large enough n, each summand of the series in (B.6) admits first the left-hand side of the inequality

$$\mathsf{P}\left\{|\xi_n(k) - \mathsf{E}\xi_n(k)| \geqslant C - (m + \tfrac{1}{2})\frac{\ln n}{n}\right\} \leqslant \mathsf{P}\{|\xi_n(k) - \mathsf{E}\xi_n(k)| \geqslant C - \varepsilon\} \quad \text{(B.18)}$$

as an upper bound, and then, of course, also the right-hand side, for all large enough n.

As shown above following (B.5), the variance of $\xi_n(k)$ is bounded from above by $c(\ln n)/n^2$ for $k > m$ and all large enough values of n. This, by Chebyshev's inequality, implies that for each such k the probability (B.7) converges to zero at the rate of $c/\ln n$, where c is some constant. For $k < m$, again, the variance of $\xi_n(k)$ is bounded from above by c'/n. By Chebyshev's inequality the probability in (B.8) converges to zero at the rate of c/n, where c is some constant. Because there are only finitely many terms in the sum in (B.6), the probability of the set D_n also converges to zero. The proof is complete.

References

AKAIKE, H. 1974 A new look at the statistical model identification. *IEEE Trans. autom. Control* AC-**19**, 716–723.

ALLEN, D. M. 1971 Mean square error of prediction as a criterion for selecting variables. *Technometrics* **13**, 469–475.

DAWID, A. P. 1984 Present position and potential developments: some personal views, statistical theory, the prequential approach. *J. R. Statist. Soc.* A **147,** 278–292.
GEISSER, S., & EDDY, W. 1979 A predictive approach to model selection. *J. Am. Statist. Assn.* **74,** 153–160.
HELMS, R. W. 1974 The average estimated variance criterion for the selection-of-variables problem in general linear regression. *Technometrics* **16,** 261–273.
HJORTH, U. 1982 Model selection for forward validation. *Scand. J. Statists* **9,** 95–105.
MALLOWS, C. L. 1973 Some Comments on C_p. *Technometrics* **15,** 661–675.
PLACKETT, R. L. 1960 *Principles of Regression Analysis.* Oxford: Clarendon Press.
PICARD, R. R., & COOK, R. D. 1984 Cross-validation of regression models. *J. Am. Statist. Assn* **79,** 575–583.
RISSANEN, J. 1978 Modeling by shortest data description. *Automatica* **14,** pp. 465–471.
RISSANEN, J. 1983 A universal prior for integers and estimation by minimum description length. *Ann. Statists* **11,** 416–431.
RISSANEN, J. 1986a Stochastic complexity and modeling. *Ann. Statists,* September 1986.
RISSANEN, J. 1986b Order estimation by accumulated prediction errors. Essays in Time Series and Allied Processes (Eds. J. Gani & M. B. Priestley). *J. appl. Probability,* Special Volume **23A,** 55–61.
STONE, M. 1974 Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc.* B **36,** 111–147.
STONE, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J.R. Statist. Soc.* B **39,** 44–47.
TOUTENBURG, H. 1982 *Prior Information in Linear Models.* Wiley.