The Annals of Statistics
1992, Vol. 20, No. 1, 1–42

ON PREDICTIVE LEAST SQUARES PRINCIPLES[1]

BY C. Z. WEI

University of Maryland and Academia Sinica

Recently, Rissanen proposed a new model selection criterion PLS that selects the model that minimizes the accumulated squares of prediction errors. Usually, the information-based criteria, such as AIC and BIC, select the model that minimizes a loss function which can be expressed as a sum of two terms. One measures the goodness of fit and the other penalizes the complexity of the selected model. In this paper we provide such an interpretation for PLS. Using this relationship, we give sufficient conditions for PLS to be strongly consistent in stochastic regression models. The asymptotic equivalence between PLS and BIC for ergodic models is then studied. Finally, based on the Fisher information, a new criterion FIC is proposed. This criterion shares most asymptotic properties with PLS while removing some of the difficulties encountered by PLS in a finite-sample situation.

1. Introduction. In this paper we are concerned with the model selection problem in regression. A large number of criteria, such as the multiple decision rule [Anderson (1963)], C_p [Mallows (1973)], AIC [Akaike (1974)], BIC [Schwarz (1978)] and cross-validation [Stone (1974)], has been proposed to solve this problem. Among them, the information-based criteria usually select the regressor $\mathbf{x}$ that minimizes the loss (or criterion) function

$$\log \hat{\sigma}_n^2 + c_n/n, \tag{1.1}$$

where n is the sample size, $\hat{\sigma}_n^2$ is the residual variance after fitting the model based on $\mathbf{x}$ and c_n is a nonnegative random variable that measures the complexity of the chosen model. For example, if $c_n = 2p$, then we have AIC, and if $c_n = p \log n$, then we have BIC. The complexity is proportional to its number of parameters.

Recently, based on his predictive minimum description length (PMDL) principle, Rissanen (1986a, b, c) proposed a new criterion that selects the regressor $\mathbf{x}$ which minimizes

$$\mathrm{PLS}(\mathbf{x}) = \sum_{i=m+1}^{n} (y_i - \mathbf{b}_{i-1}'\mathbf{x}_i)^2, \tag{1.2}$$

where y is the response variable, $\mathbf{b}_j$ is the least squares estimate based on $\{x_i, y_i\colon i \le j\}$ and m is the first integer j so that $\mathbf{b}_j$ is uniquely defined. Since $(y_i - \mathbf{b}_{i-1}'\mathbf{x}_i)^2$ is the square of the prediction error at stage i, this criterion is called the predictive least squares (PLS) principle. When the conditional

Received May 1990.
[1]Research supported by NSF Grant DMS-89-11802.
AMS 1980 *subject classifications*. Primary 62M10, 62J05; secondary 62M20.
Key words and phrases. Model selection, predictive least squares, predictive minimum description length, AIC, BIC, stochastic regression, strong consistency, FIC.

density of y_i given $\{\mathbf{x}_1, \ldots, \mathbf{x}_i, y_1, \ldots, y_{i-1}\}$ is normal, as done in Hannan (1987) for the autoregressive model, by changing the base of the logarithm and eliminating constants, one obtains

$$\text{PMDL} = \sum_{i=m+1}^{n} \left[\log \hat{\sigma}_{i-1}^2 + (y_i - \mathbf{b}_{i-1}'\mathbf{x}_i)^2/\hat{\sigma}_{i-1}^2\right]. \tag{1.3}$$

If $\lim_{n\to\infty} \hat{\sigma}_n^2 = \sigma^2$ a.s., then

$$\text{PMDL} = \left[n \log \sigma^2\right](1 + o(1)) + (\text{PLS}/\sigma^2)(1 + o(1)) \quad \text{a.s.} \tag{1.4}$$

It is clear that PLS and PMDL have a strong relationship. In this paper we concentrate on the predictive least squares principles although some previous results on PMDL will also be discussed.

There are four interrelated issues addressed in this paper. The first one deals with the following problem. Is it possible to provide PLS an interpretation as that given by (1.1)? More precisely, can one decompose PLS as a sum of a term that measures the goodness of fit and a penalty term that reflects the complexity of the model? In Section 2 we first give an identity [see (2.6)] that expresses PLS as a sum of the residual sum of squares and a penalty term. This result is natural in the sense that one expects that the accumulated error squares, due to recursive prediction, should be larger than the residual sum of squares. To give the penalty term a statistical interpretation, some asymptotic results for this term are also given (Theorems 2.2 and 2.3). Note that all the results given in this section do not require any model assumption, although the statistical meaning can be attached when a model is imposed.

The second issue deals with the strong consistency of PLS for the stochastic regression model. Rissanen (1986a, b, c) is the first one to show the weak consistency for the multiple linear regression model with Gaussian noise. Wax (1988) obtains the same result for the stationary autoregressive process without Gaussian assumption. Hannan, Mcdougall and Poskit (1989) and Hemerly and Davis (1989) independently show that Wax's result can be strengthened to be strongly consistent. The stochastic regression model [Lai and Wei (1982a); see also Section 3] not only covers multiple regression models and autoregression models, but also input–output systems that arise from the control literature. Furthermore, the consistency results described above require that for all candidate regressors $\mathbf{x}$:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' = \Gamma \quad \text{a.s.}, \tag{1.5}$$

for some positive-definite matrix Γ. This condition is violated in the optimal control systems [Lai and Wei (1986) and Davis and Hemerly (1990)] as well as in the unstable autoregressive models [Chan and Wei (1988)]. Using results from Section 2, in Section 3 we provide some sufficient conditions (Theorems 3.2 and 3.3) to obtain the strong consistency result of PLS in the stochastic regression model. Examples on fixed design models (Theorem 3.4) and unstable autoregressive processes (Theorem 3.5) are given to illustrate the general

results. The strong consistency of BIC for the unstable autoregressive model is also obtained as a side result of our analysis. For previous results on BIC, one can see Paulsen (1984) and Tsay (1984) for weak consistency and Huang (1990) for a related strong consistency result. Note that Pötscher (1989) also studies the strong consistency for the stochastic regression model. However, his result covers neither PLS nor BIC for unstable autoregressive models (see Section 5).

The third issue we are interested in is the equivalence between PLS and BIC. For a stationary AR(p_0) model,

$$y_n = \beta_1 y_{n-1} + \cdots + \beta_{p_0} y_{n-p_0} + \delta_n$$

if the order $p \geq p_0$ is selected, Hannan, Mcdougall and Poskit (1989) show that

$$\text{PLS} = n\hat{\sigma}_n^2 + \sigma^2(p \log n)(1 + o(1)) \quad \text{a.s.}, \tag{1.6}$$

where $\sigma^2 = E(\delta_n^2)$. By Taylor expansion and the fact that $\lim_{n\to\infty} \hat{\sigma}_n^2 = \sigma^2$ a.s., one has

$$\log(\text{PLS}/n) = \log \hat{\sigma}_n^2 + pn^{-1} \log n[1 + o(1)] \quad \text{a.s.} \tag{1.7}$$

Except the $o(1)$ term, this is BIC. Now it is natural to ask whether PLS is asymptotically equivalent to BIC. In this paper the asymptotical equivalence will be used in its strict sense; that is, (1.7) holds whenever the dimension of the regressor $\mathbf{x}$ is p. The result given by Hannan, Mcdougall and Poskit (1989) does not resolve the case where $p < p_0$. Recently, Kavalieris (1989) also attempts to solve this problem for the AR(∞) process. However, his result is not conclusive [see also (4.2.12) and its discussion]. Since an AR(p) process can be viewed as an AR(∞) process and for a true AR(∞) process any AR(p) fitting is misspecified, in Section 4 we study the asymptotic equivalence property when the fitted model may be incorrect. We treat the regression and time series separately. For the regression case, the regression function can be nonlinear. We obtain an asymptotic expression for PLS (Theorem 4.1) and use the polynomial regression as an example to show that PLS is not asymptotically equivalent to BIC. For the time series model, the asymptotic result for PLS is obtained under very general ergodic assumptions (Theorem 4.2.1). If the involved variables are jointly normal, then PLS is asymptotically equivalent to BIC (see Corollary 4.2.1 and the remark following the proof of Theorem 4.2.2). In particular, this is true for the Gaussian AR(∞) model.

The fourth issue is related to the performance of PLS. From our asymptotic study, it indicates that PLS is sensitive to the magnitude of the variables selected while other criteria treat each variable equally. This is a desirable feature especially for nonstationary regressors. However, as a procedure, PLS is computer intensive and tends to select the model with fewer variables when the sample size is small. Furthermore, although it is natural to use PLS for the on-line purpose, its dependency on the particular order of data seems not to be so attractive for the off-line situation. (For a detailed discussion, see Section 5.) To resolve these difficulties, a new criterion FIC, based on the

Fisher information, is proposed in Section 5. The special feature of FIC and its relationship with PLS are explained. A simulation study is also reported to demonstrate the advantage of using FIC.

Finally, some results on the unstable autoregressive process, which are used in the main text, are given in Appendix A.

2. Decomposition. Throughout this section, we assume that $\{y_i\}$ is a sequence of real numbers and $\{\mathbf{x}_i\}$ a sequence of vectors in R^p such that for some positive integer m, $V_n = (\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i')^{-1}$ exists if $n \geq m$. Let $\boldsymbol{\beta} \in R^p$. Define $\varepsilon_i = y_i - \boldsymbol{\beta}'\mathbf{x}_i$, $\mathbf{b}_n = V_n\sum_{i=1}^n \mathbf{x}_i y_i$, $\hat{\varepsilon}_i(n) = y_i - \mathbf{b}_n'\mathbf{x}_i$ and $e_i = y_i - \mathbf{b}_{i-1}'\mathbf{x}_i$.

THEOREM 2.1. *The following identity holds*:

$$\sum_{i=m+1}^n e_i^2(1 - \mathbf{x}_i'V_i\mathbf{x}_i) = \sum_{i=1}^n \hat{\varepsilon}_i^2(n) - \sum_{i=1}^m \hat{\varepsilon}_i^2(m). \tag{2.1}$$

PROOF. Let $Q_n = (\sum_{i=1}^n \mathbf{x}_i'\varepsilon_i)V_n(\sum_{i=1}^n \mathbf{x}_i\varepsilon_i)$ and $d_i = \mathbf{x}_i'V_i\mathbf{x}_i$. Then (2.8) of Wei (1987) gives

$$\begin{aligned} &Q_n - Q_m + \sum_{i=m+1}^n [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2(1 - d_i) \\ &\quad = \sum_{i=m+1}^n d_i\varepsilon_i^2 + 2\sum_{i=m+1}^n [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]\varepsilon_i(1 - d_i). \end{aligned} \tag{2.2}$$

Therefore, by this and the definition of e_i,

$$\begin{aligned} \sum_{i=m+1}^n e_i^2(1 - d_i) &= \sum_{i=m+1}^n [\varepsilon_i - \mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2(1 - d_i) \\ &= \sum_{i=m+1}^n \varepsilon_i^2(1 - d_i) - 2\sum_{i=m+1}^n [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]\varepsilon_i(1 - d_i) \\ &\quad + \sum_{i=m+1}^n [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2(1 - d_i) \\ &= \sum_{i=m+1}^n \varepsilon_i^2 - Q_n + Q_m \\ &= \left(\sum_{i=1}^n \varepsilon_i^2 - Q_n\right) - \left(\sum_{i=1}^m \varepsilon_i^2 - Q_m\right) \\ &= \sum_{i=1}^n \hat{\varepsilon}_i^2(n) - \sum_{i=1}^m \hat{\varepsilon}_i^2(m). \end{aligned}$$

□

REMARKS. (1) Note that by Lemma 2(i) of Lai and Wei (1982a),

$$\mathbf{x}_n'V_n\mathbf{x}_n = [\det(V_n^{-1}) - \det(V_{n-1}^{-1})]/\det(V_n^{-1}). \tag{2.3}$$

Since in the regression model where ε_i are i.i.d. $N(0, \sigma^2)$, the Fisher information matrix is $\sigma^{-2}V_n^{-1}$. The quantity (2.3) can be interpreted as the ratio of the information of the design point $\mathbf{x}_n$ with respect to the whole design $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

(2) Since $\det(V_n^{-1})$ increases with n, we have

$$0 \le \mathbf{x}_n' V_n \mathbf{x}_n \le 1. \tag{2.4}$$

This and (2.1) imply the intuitive fact that the sum of squares of prediction errors is larger than the residual sum of squares, that is,

$$\sum_{i=m+1}^{n} e_i^2 \ge \sum_{i=1}^{n} \hat{\varepsilon}_i^2(n) - \sum_{i=1}^{m} \hat{\varepsilon}_i^2(m). \tag{2.5}$$

This inequality is very useful when one studies the consistency property of the PLS criterion (see Section 3).

Theorem 2.1 gives a decomposition:

$$\sum_{i=m+1}^{n} e_i^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2(n) - \sum_{i=1}^{m} \hat{\varepsilon}_i^2(m) + \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i e_i^2, \tag{2.6}$$

where the first two terms can be viewed as the measure of the goodness of fit and the last one can be viewed as a penalty term. The remaining part of this section provides some asymptotic results as a first step to understand the statistical meaning of this penalty term.

THEOREM 2.2. *Assume that*

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \varepsilon_i^2 = \infty. \tag{2.7}$$

If

$$\lim_{n\to\infty} \mathbf{x}_n' V_n \mathbf{x}_n = 0 \quad \textit{and}$$

$$\sum_{i=m+1}^{n} \left[\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})\right]^2 = O\left(\sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \varepsilon_i^2 \right) \tag{2.8}$$

or

$$\lim_{n\to\infty} \mathbf{x}_n'(\mathbf{b}_{n-1} - \boldsymbol{\beta}) = 0 \quad \textit{and}$$

$$\liminf_{n\to\infty} \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \varepsilon_i^2 \Big/ \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i > 0, \tag{2.9}$$

then

$$\sum_{i=m+1}^{n} e_i^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2(n) + \left(\sum_{i=1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \varepsilon_i^2 \right)(1 + o(1)). \tag{2.10}$$

PROOF. By (2.6) and the definition of e_i,

$$\sum_{i=m+1}^{n} e_i^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2(n) - \sum_{i=1}^{m} \hat{\varepsilon}_i^2(m) + \sum_{i=m+1}^{n} d_i \varepsilon_i^2 + \sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2 d_i - 2 \sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})] \varepsilon_i d_i, \tag{2.11}$$

where $d_i = \mathbf{x}_i' V_i \mathbf{x}_i$. Hence, to show (2.10), it is sufficient to prove

$$\sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2 d_i - 2 \sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})] \varepsilon_i d_i = o\left(\sum_{i=m+1}^{n} d_i \varepsilon_i^2 \right). \tag{2.12}$$

By the Cauchy–Schwarz inequality,

$$\left| \sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})] \varepsilon_i d_i \right|^2 \le \left\{ \sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2 d_i \right\} \left\{ \sum_{i=m+1}^{n} d_i \varepsilon_i^2 \right\}.$$

Therefore, to show (2.12), we only have to prove

$$\sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2 d_i = o\left(\sum_{i=m+1}^{n} d_i \varepsilon_i^2 \right). \tag{2.13}$$

Under conditions (2.8), since $d_n = \mathbf{x}_n' V_n \mathbf{x}_n \to 0$,

$$\sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2 d_i = O(1) + o\left(\sum_{i=m+1}^{n} [\mathbf{x}_i'(b_{i-1} - \boldsymbol{\beta})]^2 \right).$$

This in turn implies (2.13) by (2.7) and (2.8). Similarly, under (2.9),

$$\sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2 = O(1) + o\left(\sum_{i=m+1}^{n} d_i \right)$$

and (2.13) follows in view of (2.7) and (2.9). □

REMARK. In the stochastic regression model [see (3.1)], (2.8) will be shown under minimal conditions. Condition (2.9) is convenient when we study the case for incorrect models (see Section 4).

The following results give the penalty term of (2.10), a further decomposition which can be interpreted statistically when the model is incorrect.

LEMMA 2.1. *Assume that there is a positive-definite matrix* Γ *such that*

$$\lim_{n \to \infty} (nV_n)^{-1} = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' = \Gamma. \tag{2.14}$$

Let δ_i be a sequence of real numbers such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \delta_i^2 = G \tag{2.15}$$

for some nonnegative-definite matrix G. Then

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \delta_i^2 = \operatorname{tr}(\Gamma^{-1} G), \tag{2.16}$$

where $\operatorname{tr}(M)$ denotes the trace of a matrix M.

PROOF. We first assume that

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=m+1}^{n} \frac{\mathbf{x}_i' \Gamma^{-1} \mathbf{x}_i \delta_i^2}{i} = \operatorname{tr}(\Gamma^{-1} G). \tag{2.17}$$

By (2.14), for any $\alpha > 0$, there is N such that for $n \geq N$, $\|nV_n - \Gamma^{-1}\| \leq \alpha/\|\Gamma\|$. Since Γ is positive definite, for all $\mathbf{x} \in R^p$, $\mathbf{x}'\Gamma^{-1}\mathbf{x} \geq \|x\|^2/\|\Gamma\|$. This in turn implies that

$$(1-\alpha)\mathbf{x}'\Gamma^{-1}\mathbf{x} \leq \mathbf{x}'(nV_n)\mathbf{x} \leq (1+\alpha)\mathbf{x}'\Gamma^{-1}\mathbf{x}, \qquad \forall\, \mathbf{x}. \tag{2.18}$$

Consequently,

$$(1-\alpha) \sum_{i=N}^{n} \frac{\mathbf{x}_i' \Gamma^{-1} \mathbf{x}_i \delta_i^2}{i} \leq \sum_{i=N}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \delta_i^2 \leq (1+\alpha) \sum_{i=N}^{n} \frac{\mathbf{x}_i' \Gamma^{-1} \mathbf{x}_i \delta_i^2}{i}. \tag{2.19}$$

The conclusion (2.16) now follows easily from (2.17) and (2.19) since α can be arbitrarily small.

Let us go back to prove (2.17). Define $S_i = \sum_{j=m+1}^{i} \mathbf{x}_j \mathbf{x}_j' \delta_j^2$. Using summation by parts and the convention $S_m = 0$,

$$\begin{aligned}
\sum_{i=m+1}^{n} \frac{\mathbf{x}_i' \Gamma^{-1} \mathbf{x}_i \delta_i^2}{i} &= \sum_{i=m+1}^{n} \frac{\operatorname{tr}\left(\Gamma^{-1} \mathbf{x}_i \mathbf{x}_i' \delta_i^2\right)}{i} \\
&= \sum_{i=m+1}^{n} \frac{\operatorname{tr}\left[\Gamma^{-1}(S_i - S_{i-1})\right]}{i} \\
&= \frac{\operatorname{tr}\left(\Gamma^{-1} S_n\right)}{n} + \sum_{i=m+1}^{n-1} \operatorname{tr}\left(\Gamma^{-1} S_i\right)\left[i^{-1} - (i+1)^{-1}\right] \\
&= \operatorname{tr}\left(\frac{\Gamma^{-1} S_n}{n}\right) + \sum_{i=m+1}^{n-1} \operatorname{tr}\left(\frac{\Gamma^{-1} S_i}{i}\right) \frac{1}{i+1}.
\end{aligned} \tag{2.20}$$

By (2.15) and (2.20), (2.17) is proved. □

THEOREM 2.3. *Assume that* (2.14) *holds and* $\varepsilon_i = h_i + \alpha_i$. *If there exist a nonnegative definite matrix* $\tilde{G}$ *and a nonnegative number* σ^2 *such that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' h_i^2 = \tilde{G}, \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \alpha_i^2 = \sigma^2 \Gamma \tag{2.21}$$

and

$$\sum_{i=m+1}^{n} \mathbf{x}_i V_i \mathbf{x}_i h_i \alpha_i = O\left(\sum_{i=m+1}^{n} (\mathbf{x}_i' V_i \mathbf{x}_i)^2 h_i^2 \right), \tag{2.22}$$

then

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \varepsilon_i^2 = p\sigma^2 + \operatorname{tr}(\Gamma^{-1}\tilde{G}). \tag{2.23}$$

If we assume furthermore that

$$\lim_{n\to\infty} \mathbf{x}_n'(\mathbf{b}_{n-1} - \boldsymbol{\beta}) = 0, \tag{2.24}$$

then

$$\sum_{i=m+1}^{n} e_i^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2(n) + (\log n)\left[p\sigma^2 + \operatorname{tr}(\Gamma^{-1}\tilde{G}) \right](1 + o(1)). \tag{2.25}$$

PROOF. First observe that

$$\begin{aligned} \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \varepsilon_i^2 = {} & \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i h_i^2 + 2 \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i h_i \alpha_i \\ & + \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \alpha_i^2. \end{aligned} \tag{2.26}$$

By Lemma 2.1 and (2.21),

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=m+1}^{n} \mathbf{x}' V_i \mathbf{x}_i h_i^2 = \operatorname{tr}(\Gamma^{-1}\tilde{G}) \tag{2.27}$$

and

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \alpha_i^2 = \sigma^2 \operatorname{tr}(\Gamma^{-1}\Gamma) = p\sigma^2. \tag{2.28}$$

Now by (2.14), $\mathbf{x}_n' V_n \mathbf{x}_n \to 0$. This, (2.22) and (2.27) imply that

$$\sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i h_i \alpha_i = o\left(\sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i h_i^2 \right) + O(1) = o(\log n). \tag{2.29}$$

Consequently, (2.23) follows from (2.26)–(2.29).

Let us show (2.25) by using Theorem 2.2. First, (2.14) and Lemma 2.1 with $\delta_i = 1$ imply that

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i = \operatorname{tr}(\Gamma^{-1}\Gamma) = p. \tag{2.30}$$

In view of (2.23), (2.30) and (2.24), (2.7) and (2.9) hold. Therefore, (2.10) holds. Combining this and (2.23), we obtain (2.25). □

REMARKS. (1) In applications (see Section 4), h_i is the bias due to fitting and α_i is the random error which in general is independent of $\{\mathbf{x}_1, \ldots, \mathbf{x}_i, h_{m+1}, \ldots, h_i\}$. Under minimal assumptions (see the proof of Theorem 4.1), (2.22) is a consequence of Chow's (1965) local martingale convergence result.

(2) In the conventional criterion, the penalty term increases linearly with p. However, in (2.25) the second term $\operatorname{tr}(\Gamma^{-1}\tilde{G})$, which in general can be viewed as a standardized measure of the goodness of fit, is expected to decrease as p increases.

3. Consistency. In this section we consider the following stochastic regression model:

$$y_n = \boldsymbol{\beta}'\mathbf{x}_n + \delta_n, \tag{3.1}$$

where $\{\delta_n\}$ is a sequence of martingale differences with respect to σ-fields $\{\mathscr{F}_n\}$ and $\mathbf{x}_n$ is an $\mathscr{F}_{n-1}$-measurable, p-dimensional random vector. Fixed design regression models, autoregressive time series and linear input–output control systems are some important stochastic regression models.

To study the effect on the PLS by adding (or deleting) variables, we need some notation. Let $T_n = (x_{n2}, \ldots, x_{np})'$ and rewrite $V_n^{-1} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'$ as

$$V_n^{-1} = \begin{pmatrix} \sum_{i=1}^n x_{il}^2 & K_n \\ K_n' & H_n \end{pmatrix}.$$

Define $\delta_i(n) = y_i - T_i'H_n^{-1}\sum_{j=1}^n T_j y_j$ and $s_n^2 = \sum_{i=1}^n (x_{i1} - K_n H_n^{-1} T_i)^2$. Note that s_n^2 is the residual sum of squares obtained by regressing x_1 on $x_2, \ldots, x_p$. The following result gives a lower bound for the difference between the residual sum of squares obtained by regressing y on $x_2, \ldots, x_p$ and $\sum_{i=1}^n \delta_i^2$.

THEOREM 3.1. *Assume that* (3.1) *holds. If for some* $\alpha > 2$,

$$\sup_n E\left[|\delta_n|^\alpha \mathscr{F}_{n-1}\right] < \infty \quad a.s., \tag{3.2}$$

and

$$s_n^2 \to \infty, \qquad \log\left(\sum_{i=1}^n \|T_i\|^2\right) = o(s_n^2) \quad a.s., \tag{3.3}$$

then

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n \delta_i^2(n) - \sum_{i=1}^n \delta_i^2}{s_n^2} = \beta_1^2 \quad a.s. \tag{3.4}$$

PROOF. Let L be the linear space generated by

$$\{X_{jn} = (x_{1j}, \ldots, x_{nj})' : 2 \le j \le p\}.$$

Denote the projection of the vector U onto L by U^*. Let $Z_n = \beta_1 X_{1n}$ and $\boldsymbol{\delta}_n = (\delta_1, \ldots, \delta_n)'$. Then

$$\begin{aligned} \sum_{i=1}^{n} \delta_i^2(n) - \sum_{i=1}^{n} \delta_i^2 &= \|Z_n - Z_n^* + \boldsymbol{\delta}_n - \boldsymbol{\delta}_n^*\|^2 - \|\boldsymbol{\delta}_n\|^2 \\ &= \|Z_n - Z_n^*\|^2 + 2\langle Z_n - Z_n^*, \boldsymbol{\delta}_n\rangle - \|\boldsymbol{\delta}_n^*\|^2 \\ &= \beta_1^2 s_n^2 + 2\beta_1 \sum_{i=1}^{n} (x_{i1} - K_n H_n^{-1} T_i)\delta_i - \|\boldsymbol{\delta}_n^*\|^2. \end{aligned} \tag{3.5}$$

By Theorem 3(i) and Lemma 3 of Lai and Wei (1982b),

$$\|\boldsymbol{\delta}_n^*\|^2 = O\left(\left\{\log \sum_{i=1}^{n} \|T_i\|^2\right\}\right) \quad \text{a.s.}, \tag{3.6}$$

and

$$\sum_{i=1}^{n} (x_{i1} - K_n H_n^{-1} T_i)\delta_i = O\left(s_n\left\{\log s_n^2 + \log \sum_{i=1}^{n} \|T_i\|^2\right\}\right) \quad \text{a.s.} \tag{3.7}$$

In view of (3.5)–(3.7) and (3.3), (3.4) follows. □

REMARK. Let $\lambda^*(M)$ and $\lambda_*(M)$ denote the maximum and minimum eigenvalues of a matrix M. We then have

$$\sum_{i=1}^{n} \|T_i\|^2 \le \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 = \operatorname{tr}(V_n^{-1}) \le p\lambda^*(V_n^{-1}) \tag{3.8}$$

and, by (1.6) of Lai and Wei (1982b),

$$s_n^2 \ge p^{-1}\lambda_*(V_n^{-1}). \tag{3.9}$$

Therefore, (3.3) is a corollary of the assumption

$$\lambda_*(V_n^{-1}) \to \infty \quad \text{and} \quad \log \lambda^*(V_n^{-1}) = o(\lambda_*(V_n^{-1})) \quad \text{a.s.} \tag{3.10}$$

A weaker version of (3.4) is proved in Pötscher (1989) under assumption (3.10).

In the following, the least squares estimate of $\boldsymbol{\beta}$ in (3.1) is denoted by $\mathbf{b}_n$ and $\hat{\delta}_i(n) = y_i - \mathbf{x}_i'\mathbf{b}_n$.

LEMMA 3.1. *Assume that* (3.1) *and* (3.2) *hold. If*

$$\limsup_{n\to\infty} \mathbf{x}_n' V_n \mathbf{x}_n < 1 \quad a.s., \tag{3.11}$$

then

$$\sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2 = O\left(1 + \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \varepsilon_i^2\right) \tag{3.12}$$

$$= O\left(\{\log \det(V_n^{-1})\}^{1+\delta}\right) \quad a.s.,$$

where $m = \inf\{j\colon V_j^{-1}$ *exists*$\}$.

PROOF. By (2.17) and (2.18) of Lai and Wei (1982a),

$$\left\{\sum_{i=m+1}^{n} [\mathbf{x}_i'(\mathbf{b}_{i-1} - \boldsymbol{\beta})]^2 (1 - \mathbf{x}_i' V_i \mathbf{x}_i)\right\}(1 + o(1))$$

$$= O(1) + \sum_{i=m+1}^{n} \mathbf{x}_i' V_i \mathbf{x}_i \varepsilon_i^2 \quad \text{a.s.}$$

By (3.11), the first identity of (3.12) is obtained. The second identity follows from (2.21) of Lai and Wei (1982a). □

Let $C_1(n)$ and $C_2(n)$ denote the PLS by selecting a subset $M \subset \{x_2, \ldots, x_p\}$ and $\{x_1, \ldots, x_p\}$, respectively.

THEOREM 3.2. *Assume that* (3.1), (3.2), (3.10) *and* (3.11) *hold. Then*

$$\liminf_{n\to\infty} \frac{C_1(n) - C_2(n)}{s_n^2} \geq \beta_1^2 \quad a.s. \tag{3.13}$$

PROOF. Note that the residual sum of squares based on M is larger than the residual sum of squares based on $\{x_2, \ldots, x_p\}$. By Theorem 2.1 and Chow's (1965) result,

$$C_1(n) - C_2(n) + O(1) \geq \sum_{i=1}^{n} \delta_i^2(n) - \sum_{i=m+1}^{n} [y_i - \mathbf{x}_i'\mathbf{b}_{i-1}]^2$$

$$= \left[\sum_{i=1}^{n} \delta_i^2(n) - \sum_{i=1}^{n} \delta_i^2\right] - 2\sum_{i=m+1}^{n} \mathbf{x}_i'(\boldsymbol{\beta} - \mathbf{b}_{i-1})\delta_i$$

$$- \sum_{i=m+1}^{n} [\mathbf{x}_i'(\boldsymbol{\beta} - \mathbf{b}_{i-1})]^2 \tag{3.14}$$

$$= \left[\sum_{i=1}^{n} \delta_i^2(n) - \sum_{i=1}^{n} \delta_i^2\right]$$

$$- \left\{\sum_{i=m+1}^{n} [\mathbf{x}_i'(\boldsymbol{\beta} - \mathbf{b}_{i-1})]^2\right\}(1 + o(1)) + O(1) \quad \text{a.s.}$$

In view of Theorem 3.1, Lemma 3.1, (3.9), (3.10) and (3.14), (3.13) follows. □

Given a set of variables $\mathbf{x}$, we define it as a correct model if (3.1) holds for some β and δ. Assume that there exists a correct model, say $\mathbf{x}$. Then for any incorrect model, we can always add variables from $\mathbf{x}$, so that the enlarged model is correct. Theorem 3.2 provides conditions which ensure that this incorrect model would not be chosen by the predictive least squares principle eventually. To find conditions that ensure the selection of the "desired" one (say with least variables or least order), more delicate study of the terms $\sum_{i=m+1}^{n}\mathbf{x}_i'V_i\mathbf{x}_i\delta_i^2$ and $Q_n = (\sum_{i=1}^{n}\mathbf{x}_i'\delta_i)V_n(\sum_{i=1}^{n}\mathbf{x}_i\delta_i) = \sum_{i=1}^{n}\delta_i^2 - \sum_{i=1}^{n}(y_i - \mathbf{b}_n'\mathbf{x}_i)^2$ is required.

LEMMA 3.2. *Assume that* $\{\delta_n, \mathcal{F}_n\}$ *is a sequence of martingale differences such that for some* $\sigma^2 > 0$ *and* $\alpha > 2$,

$$E(\delta_n^2|\mathcal{F}_{n-1}) = \sigma^2 \quad \text{and} \quad \sup_n E(|\delta_n|^\alpha|\mathcal{F}_{n-1}) < \infty \quad a.s. \tag{3.15}$$

If $\mathbf{x}_n$ *is* $\mathcal{F}_{n-1}$*-measurable,* $\lambda_*(V_n^{-1}) \to \infty$ *a.s. and*

$$\mathbf{x}_n'V_n\mathbf{x}_n \to 0 \quad a.s., \tag{3.16}$$

then

$$\sum_{i=m}^{n}\mathbf{x}_i'V_i\mathbf{x}_i\delta_i^2 \sim \sigma^2\log\det(V_n^{-1}) \quad a.s. \tag{3.17}$$

This lemma is shown in Wei (1987), (2.12) and (2.14). □

LEMMA 3.3. *The following identities hold*:

$$\mathbf{x}_n'V_n\mathbf{x}_n = \frac{s_n^2 - s_{n-1}^2}{s_n^2} + \frac{s_{n-1}^2}{s_n^2}T_n'H_n^{-1}T_n, \tag{3.18}$$

$$\log\det(V_n^{-1}) = \log(s_n^2) + \log\det(H_n), \tag{3.19}$$

$$\begin{aligned}\left(\sum_{i=1}^{n}\mathbf{x}_i'\delta_i\right)V_n\left(\sum_{i=1}^{n}\mathbf{x}_i\delta_i\right) &= \left[\sum_{i=1}^{n}(x_{i1} - K_nH_n^{-1}T_i)\delta_i\right]^2\frac{1}{s_n^2} \\ &\quad + \left(\sum_{i=1}^{n}T_i'\delta_i\right)H_n^{-1}\left(\sum_{i=1}^{n}T_i\delta_i\right).\end{aligned} \tag{3.20}$$

PROOF. By Lemma 4(vi) of Lai, Robbins and Wei (1979),

$$1 + \mathbf{x}_n'V_{n-1}\mathbf{x}_n = (s_n^2/s_{n-1}^2)(1 + T_n'H_{n-1}^{-1}T_n)$$

or

$$(1 - \mathbf{x}_n'V_n\mathbf{x}_n)^{-1} = (s_n^2/s_{n-1}^2)(1 - T_n'H_nT_n)^{-1}.$$

By simple algebra, we obtain (3.18). For (3.19), observe that $X_{1n}^* = (x_{11} - K_nH_n^{-1}T_1, \ldots, x_{n1} - K_nH_n^{-1}T_n)'$ is orthogonal to $X_{in} = (x_{1i}, \ldots, x_{ni})'$ for $i \geq 2$. Let

$$W_n = \begin{pmatrix} 1 & -K_nH_n^{-1} \\ \mathbf{0} & I \end{pmatrix},$$

where I is the $(p-1)\times(p-1)$ identity matrix and $\mathbf{0}$ is the $(p-1)$ zero vector. We then have

$$W_n(V_n^{-1})W_n' = \sum_{i=1}^{n}(W_n\mathbf{x}_i)(W_n\mathbf{x}_i)' = \begin{pmatrix} s_n^2 & \mathbf{U}_n \\ \mathbf{U}_n' & H_n \end{pmatrix}, \tag{3.21}$$

where

$$\mathbf{U}_n = \sum_{i=1}^{n}(x_{i1} - K_nH_n^{-1}T_i)T_i = (X_{2n}'X_{1n}^*, \ldots, X_{pn}'X_{1n}^*)' = \mathbf{0}.$$

Thus

$$\begin{aligned}\det(V_n^{-1}) &= \det(W_n)\det(V_n^{-1})\det(W_n') \\ &= \det(W_nV_n^{-1}W_n') = s_n^2\det(H_n)\end{aligned}$$

and (3.19) follows. For (3.20), it is an immediate consequence of (3.21) and

$$\left(\sum_{i=1}^{n}\mathbf{x}_i'\delta_i\right)V_n\left(\sum_{i=1}^{n}\mathbf{x}_i\delta_i\right) = \left(\sum_{i=1}^{n}W_n\mathbf{x}_i\delta_i\right)'\left[W_nV_n^{-1}W_n'\right]^{-1}\left(\sum_{i=1}^{n}W_n\mathbf{x}_i\delta_i\right). \quad \square$$

LEMMA 3.4. *Assume that the martingale difference sequence* $\{\delta_n, \mathcal{F}_n\}$ *satisfies*

$$\sup_n E\left(|\delta_n|^\alpha \middle| \mathcal{F}_{n-1}\right) < \infty \quad a.s., \tag{3.22}$$

for some $\alpha > 2$ *and* $\lambda_*(V_n^{-1}) \to \infty$.

If $\mathbf{x}_n$ *are nonrandom vectors, then*

$$\left(\sum_{i=1}^{n}\mathbf{x}_i\delta_i\right)V_n\left(\sum_{i=1}^{n}\mathbf{x}_i\delta_i\right) = o\left(\log\lambda^*(V_n^{-1})\right) \quad a.s. \tag{3.23}$$

PROOF. Let $Q_n = (\sum_{i=1}^n\mathbf{x}_i\delta_i)V_n(\sum_{i=1}^n\mathbf{x}_i\delta_i)$. Consider the case $p=1$. Since $\sum_{i=1}^\infty\|\mathbf{x}_i\|^2 = \infty$, by (2.30) of Wei (1987), for every $\gamma > 2\alpha^{-1}$,

$$Q_n = o\left(\left[\log\left(\sum_{i=1}^{n}\|\mathbf{x}_i\|^2\right)\right]^\gamma\right) \quad \text{a.s.}$$

Using the fact that $\alpha > 2$,

$$Q_n = o\left(\log\left(\sum_{i=1}^{n}\|\mathbf{x}_i\|^2\right)\right) = o\left(\log\lambda^*(V_n^{-1})\right).$$

Now, let us consider the case $p > 1$. By (3.20) of Lemma 3.3, an induction argument and the facts that $\lambda^*(V_n^{-1}) \geq \lambda^*(H_n)$ and $\lambda_*(V_n^{-1}) \leq \lambda_*(H_n)$, it is sufficient to show that

$$\left[\sum_{i=1}^{n}(x_{i1} - K_nH_n^{-1}T_i)\delta_i\right]^2\frac{1}{s_n^2} = o\left(\log\lambda^*(V_n^{-1})\right) \quad \text{a.s.} \tag{3.24}$$

First, let us assume that there is $K > 0$ such that

$$\sup_n E\left(|\delta_n|^\alpha \middle| \mathscr{F}_{n-1}\right) \le K \quad \text{a.s.} \tag{3.25}$$

In view of (3.9), $s_n^2 \to \infty$. This and Corollary 2 of Lai and Wei (1984) imply that for all $\gamma > 2\alpha^{-1}$,

$$\left[\sum_{i=1}^n \left(X_{i1} - K_n H_n^{-1} T_i\right)\delta_i\right]^2 \frac{1}{s_n^2} = o\left(\left(\log s_n^2\right)^\gamma\right) \quad \text{a.s.}$$

The conclusion (3.23) follows from the fact that $\alpha > 1/2$, $s_n^2 \le \lambda^*(V_n^{-1})$, (3.20) and an induction argument. Finally, let us remove condition (3.25). By the above arguments, (3.23) holds if we replace δ_n by $\delta_n I_{[E(|\delta_n|^\alpha|\mathscr{F}_{n-1}) \le K]}$. Therefore, it holds on the set

$$\left\{E\left(|\delta_n|^\alpha \middle| \mathscr{F}_{n-1}\right) \le K, \forall\, n\right\} \supseteq \left\{\sup_n E\left(|\varepsilon_n|^\alpha \middle| \mathscr{F}_{n-1}\right) \le K\right\}.$$

Letting $K \to \infty$, we complete our proof. □

COUNTEREXAMPLE. When $\mathbf{x}_n$ are random vectors, (3.23) may not hold. For this, consider the following example [Lai and Wei (1982a)]:

$$y_i = \beta_1 + \beta_2 x_i + \delta_i, \tag{3.26}$$

where δ_i are i.i.d. random variables with $E(\delta_i) = 0$, $E(\delta_i^2) = 1$ and x_i are defined inductively by

$$x_1 = 0, \qquad x_{n+1} = \bar{x}_n + c\bar{\delta}_n, \qquad n \ge 1,$$

where $c \ne 0$. It is known [Lai and Wei (1982a), page 160] that

$$b_{n2} - \beta_2 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)\delta_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \to -\frac{1}{c} \quad \text{a.s.}, \tag{3.27}$$

$$\sum_{i=1}^n (x_i - \bar{x}_n)^2 \sim c^2 \log n \quad \text{a.s.}, \tag{3.28}$$

$$\lambda^*\left(V_n^{-1}\right) \sim n\left\{1 + c^2\left(\sum_{j=1}^\infty \frac{\varepsilon_j}{j}\right)\right\} \quad \text{a.s.} \tag{3.29}$$

Now by (3.20), (3.27)–(3.29) and the law of iterated logarithm,

$$\begin{aligned} Q_n &= \frac{\left\{\sum_{i=1}^n (x_i - \bar{x}_n)\delta_i\right\}^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} + \frac{\left(\sum_{i=1}^n \delta_i\right)^2}{n} \\ &= (b_{n2} - \beta_2)^2\left[\sum_{i=1}^n (x_i - \bar{x}_n)^2\right] + O(\log\log n) \\ &\sim \log n + O(\log\log n) \\ &\sim \log \lambda^*\left(V_n^{-1}\right) \quad \text{a.s.} \end{aligned}$$

Therefore, (3.23) is violated.

The above counterexample shows that for the stochastic regression model, it is impossible to obtain (3.23) without further conditions. The following results provide a remedy.

LEMMA 3.5. *Assume that* (3.1) *and* (3.22) *hold. Also assume that there exists a nonsingular matrix A such that* $A\mathbf{x}_n = (\mathbf{z}_n', \mathbf{w}_n')$ *satisfies*

$$\left(\sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i'\right)^{-1/2}\left(\sum_{i=1}^{n} \mathbf{z}_i\mathbf{w}_i'\right)\left(\sum_{i=1}^{n} \mathbf{w}_i\mathbf{w}_i'\right)^{-1/2} \to 0 \quad a.s., \tag{3.30}$$

and

$$\liminf_{n\to\infty} \lambda_*\left(D_n^{-1}\left(\sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i'\right)D_n^{-1}\right) > 0 \quad a.s., \tag{3.31}$$

where $D_n = \{\operatorname{diag}(\sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i')\}^{1/2}$.

Then

$$\begin{aligned}&\left(\sum_{i=1}^{n} \mathbf{x}_i'\delta_i\right)V_n\left(\sum_{i=1}^{n} \mathbf{x}_i\delta_i\right)\\ &= \left(\sum_{i=1}^{n} \mathbf{w}_i'\delta_i\right)\left(\sum_{i=1}^{n} \mathbf{w}_i\mathbf{w}_i'\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{w}_i\delta_i\right) + o\left(\log \lambda^*\left(V_n^{-1}\right)\right) \quad a.s.\end{aligned} \tag{3.32}$$

PROOF. Let $\mathbf{u}_n = (\mathbf{z}_n^*, \mathbf{w}_n')'$, $J_n = \sum_{i=1}^{n} \mathbf{u}_i'\mathbf{u}_i$, $P_n = \sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i'$ and $G_n = \sum_{i=1}^{n} \mathbf{w}_i\mathbf{w}_i'$. Since A is nonsingular,

$$\begin{aligned}Q_n &= \left(\sum_{i=1}^{n} \mathbf{x}_i'\delta_i\right)V_n\left(\sum_{i=1}^{n} \mathbf{x}_i\delta_i\right)\\ &= \left(\sum_{i=1}^{n} A\mathbf{x}_i\delta_i\right)'\left[\sum_{i=1}^{n} A\mathbf{x}_i(A\mathbf{x}_i)'\right]^{-1}\left(\sum_{i=1}^{n} A\mathbf{x}_i\delta_i\right)\\ &= \left(\sum_{i=1}^{n} \mathbf{u}_i'\delta_i\right)J_n^{-1}\left(\sum_{i=1}^{n} \mathbf{u}_i\delta_i\right).\end{aligned} \tag{3.33}$$

By (3.30),

$$\begin{pmatrix} P_n^{-1/2} & 0 \\ 0 & G_n^{-1/2} \end{pmatrix} J_n \begin{pmatrix} P_n^{-1/2} & 0 \\ 0 & G_n^{-1/2} \end{pmatrix} \to I \quad \text{a.s.}$$

This in turn implies that

$$Q_n = \left(\left\| P_n^{-1/2} \sum_{i=1}^{n} \mathbf{z}_i \delta_i \right\|^2 + \left\| G_n^{-1/2} \sum_{i=1}^{n} \mathbf{w}_i \delta_i \right\|^2 \right)(1 + o(1))$$

$$= \left[\left(\sum_{i=1}^{n} \mathbf{z}_i' \delta_i \right) P_n^{-1} \left(\sum_{i=1}^{n} \mathbf{z}_i \delta_i \right) \right. \tag{3.34}$$

$$\left. + \left(\sum_{i=1}^{n} \mathbf{w}_i' \delta_i \right) G_n^{-1} \left(\sum_{i=1}^{n} \mathbf{w}_i \delta_i \right) \right] (1 + o(1)).$$

Since $\lambda_*(V_n^{-1}) \to \infty$, $\lambda_*(P_n) \to \infty$. In view of this, (3.31) and (2.33) of Wei (1987), for $\gamma > 2\alpha^{-1}$,

$$\left(\sum_{i=1}^{n} \mathbf{z}_i' \delta_i \right) P_n^{-1} \left(\sum_{i=1}^{n} \mathbf{z}_i \delta_i \right) = o\left(\left(\log \sum_{i=1}^{n} \| \mathbf{z}_i \| \right)^{\gamma} \right) \tag{3.35}$$

$$= o\left(\log \lambda^*\left(V_n^{-1} \right) \right) \quad \text{a.s.}$$

Combining (3.34), (3.35) and the fact [Lai and Wei (1982a)] that $Q_n = O(\log \lambda^*(V_n^{-1}))$, (3.32) follows. □

REMARK. If $A\mathbf{x}_n = \mathbf{z}_n$ and $\mathbf{z}_n$ satisfies (3.31), then (3.30) can be omitted and (3.32) holds without the first term. The same remark also applies to the following corollary.

COROLLARY 3.1. *Assume that* (3.1), (3.22), (3.30) *and* (3.31) *hold. If* $\beta_1 = 0$ *and* A *can be expressed as*

$$A = \begin{pmatrix} a_{11} & \mathbf{a}' \\ \mathbf{0} & B \end{pmatrix}, \tag{3.36}$$

where $\mathbf{a} \in R^{p-1}$ *and* B *is a* $(p-1) \times (p-1)$ *nonsingular matrix, then*

$$\left(\sum_{i=1}^{n} \mathbf{x}_i' \delta_i \right) V_n \left(\sum_{i=1}^{n} \mathbf{x}_i \delta_i \right) - \left(\sum_{i=1}^{n} T_i' \delta_i \right) H_n^{-1} \left(\sum_{i=1}^{n} T_i \delta_i \right) \tag{3.37}$$

$$= o\left(\log \lambda^*\left(V_n^{-1} \right) \right) \quad a.s.$$

PROOF. Let $\mathbf{z}_n' = (z_{n1}, \ldots, z_{nq})$ and $\mathbf{t}_n' = (z_{n2}, \ldots, z_{nq})$. Then, by (3.36), $BT_n = (\mathbf{t}_n', \mathbf{w}_n')'$. By (3.31), it is easy to see that

$$\liminf_{n \to \infty} \lambda_* \left(\hat{D}_n^{-1} \left(\sum_{i=1}^{n} \mathbf{t}_i \mathbf{t}_i' \right) \hat{D}_n^{-1} \right) > 0 \quad \text{a.s.}, \tag{3.38}$$

where $\hat{D}_n = \{\text{diag}(\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i')\}^{1/2}$.

Also observe that under (3.31), (3.30) holds if and only if

$$D_n^{-1/2} \left(\sum_{i=1}^{n} \mathbf{z}_i \mathbf{w}_i' \right) \left(\sum_{i=1}^{n} \mathbf{w}_i \mathbf{w}_i' \right)^{-1/2} \to 0 \quad \text{a.s.} \tag{3.39}$$

But (3.39) implies that

$$\hat{D}_n^{-1/2}\left(\sum_{i=1}^{n}\mathbf{t}_i\mathbf{w}_i'\right)\left(\sum_{i=1}^{n}\mathbf{w}_i\mathbf{w}_i'\right)^{-1/2} \to 0 \quad \text{a.s.}$$

Therefore, by (3.38),

$$\left(\sum_{i=1}^{n}\mathbf{t}_i\mathbf{t}_i'\right)^{-1/2}\left(\sum_{i=1}^{n}\mathbf{t}_i\mathbf{w}_i'\right)\left(\sum_{i=1}^{n}\mathbf{w}_i\mathbf{w}_i'\right)^{-1/2} \to 0 \quad \text{a.s.} \tag{3.40}$$

Replacing $\mathbf{x}_i$ in Lemma 3.5 by T_i and A by B, (3.38) and (3.40) give

$$\begin{aligned}&\left(\sum_{i=1}^{n}T_i'\delta_i\right)H_n^{-1}\left(\sum_{i=1}^{n}T_i\delta_i\right)\\ &\quad=\left(\sum_{i=1}^{n}\mathbf{w}_i'\delta_i\right)\left(\sum_{i=1}^{n}\mathbf{w}_i\mathbf{w}_i'\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{w}_i\delta_i\right)+o\left(\log\lambda^*\left(V_n^{-1}\right)\right)\quad\text{a.s.}\end{aligned} \tag{3.41}$$

Combining (3.32) and (3.41), we complete the proof of (3.37). □

Now we are ready to state a result that provides conditions under which the correct model with least variables would be chosen eventually. Let $C_1(n)$ and $C_2(n)$ denote the PLS by selecting $\{x_2,\ldots,x_p\}$ and $\{x_1,\ldots,x_p\}$, respectively. Note that if $\beta_1 = 0$, both models are correct.

THEOREM 3.3. *Assume that* (3.1), (3.15), (3.16) *and* (3.22) *hold. Also assume that* $\beta_1 = 0$ *and*

$$\liminf_{n\to\infty}\frac{\log\lambda^*\left(V_n^{-1}\right)}{\log\lambda_*\left(V_n^{-1}\right)} > 0 \quad a.s. \tag{3.42}$$

If either $\mathbf{x}_n$ *are nonrandom vectors or* (3.30), (3.31) *and* (3.36) *are satisfied, then*

$$\lim_{n\to\infty}\frac{C_2(n)-C_1(n)}{\log s_n^2} = \sigma^2 \quad a.s. \tag{3.43}$$

PROOF. Let

$$Q_n = \left(\sum_{i=1}^{n}\mathbf{x}_i'\delta_i\right)V_n\left(\sum_{i=1}^{n}\mathbf{x}_i\delta_i\right) \quad \text{and} \quad \hat{Q}_n = \left(\sum_{i=1}^{n}T_i'\delta_i\right)H_n^{-1}\left(\sum_{i=1}^{n}T_i\delta_i\right).$$

By Lemmas 3.1 and 3.2 and Theorem 2.2,

$$C_2(n) = \sum_{i=1}^{n}\delta_i^2 - Q_n + \left[\sigma^2\log\det\left(V_n^{-1}\right)\right](1+o(1)) \quad \text{a.s.,} \tag{3.44}$$

and

$$C_1(n) = \sum_{i=1}^{n}\delta_i^2 - \hat{Q}_n + \left[\sigma^2\log\det\left(H_n\right)\right](1+o(1)) \quad \text{a.s.} \tag{3.45}$$

Therefore,

$$C_2(n) - C_1(n) = \sigma^2\left[\log\det\left(V_n^{-1}\right) - \log\det(H_n)\right] - \left[Q_n - \hat{Q}_n\right] + o\left(\log\det\left(V_n^{-1}\right)\right) \quad \text{a.s.} \tag{3.46}$$

By (3.19) of Lemma 3.3,

$$\sigma^2\left[\log\det\left(V_n^{-1}\right) - \log\det(H_n)\right] = \left(\log s_n^2\right)\sigma^2. \tag{3.47}$$

Now, if $\mathbf{x}_n$ are nonrandom vectors, applying Lemma 3.4 on $\mathbf{x}_n$ and T_n, we have

$$Q_n - \hat{Q}_n = o\left(\log\det\left(V_n^{-1}\right)\right) \text{ a.s.} \tag{3.48}$$

If $\mathbf{x}_n$ are random vectors, (3.48) also holds through Lemma 3.5. Combining (3.46)–(3.48), (3.42) and (3.9), (3.43) is obtained. □

To illustrate the applications of Theorems 3.2 and 3.3, we consider two examples.

EXAMPLE 1 (Fixed design). Let $M^* = \{x_1, \ldots, x_p\}$ be the set of all independent variables to be chosen. Based on the observations $\{y_1, \ldots, y_n, x_{1j}, \ldots, x_{nj}, 1 \le j \le p\}$, we would like to select a model $M \in \mathcal{M} = \{M\colon M \subset M^*\}$ so that M is a correct model with least variables. Let $M_0 = \{x_{l_1}, \ldots, x_{l_q}\}$, $\mathbf{x}_i^0 = (x_{il_1}, \ldots, x_{il_q})'$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ and

$$y_i = \boldsymbol{\gamma}'\mathbf{x}_i^0 + \delta_i, \tag{3.49}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$. We assume that $\{\delta_n\}$ are i.i.d. random variables with $E(\delta_n) = 0$, $E(\delta_n^2) = \sigma^2 > 0$ and $E|\delta_n|^\alpha < \infty$ for some $\alpha > 2$. We also assume that $\gamma_i \neq 0$, $i = 1, \ldots, q$. It is not difficult to see that if $V_n = (\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i')^{-1}$ exists, M_0 is the unique model with least variables. In the following, $\hat{M}_n$ denotes the model chosen by the predictive least squares principle.

THEOREM 3.4. *Assume that there exists a sequence of positive real numbers a_n such that*

$$\lim_{n\to\infty} a_n^{-1}a_{n+1} = 1 \tag{3.50}$$

and for some positive-definite matrix Γ,

$$\lim_{n\to\infty} \frac{1}{a_n}\left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right) = \Gamma. \tag{3.51}$$

Then

$$P\left[\hat{M}_n = M_0 \text{ eventually}\right] = 1. \tag{3.52}$$

PROOF. Note that

$$\begin{aligned}\mathbf{x}_n' V_n \mathbf{x}_n &= \operatorname{tr}(V_n \mathbf{x}_n \mathbf{x}_n') = \operatorname{tr}\left[V_n\left(V_n^{-1} - V_{n-1}^{-1}\right)\right] \\ &= \operatorname{tr}\left[I - a_n^{-1} a_{n-1}(a_n V_n)\left(a_{n-1}^{-1} V_{n-1}^{-1}\right)\right] \\ &\to \operatorname{tr}[I - I] = 0\end{aligned} \tag{3.53}$$

by (3.50) and (3.51). Furthermore, (3.51) also implies that

$$\lim_{n\to\infty} \lambda^*\left(V_n^{-1}\right)/\lambda_*\left(V_n^{-1}\right) > 0. \tag{3.54}$$

Given any subvector $\mathbf{u}_i$ of $\mathbf{x}_i$, (3.51) also implies that

$$\lim_{n\to\infty} \frac{1}{a_n}\left(\sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^*\right) = \Gamma^*,$$

where Γ^* is also positive definite. Consequently, (3.53) and (3.54) also hold if $\mathbf{x}_i$ are replaced by $\mathbf{u}_i$. Thus Theorems 3.2 and 3.3 are applicable if $\mathbf{u}_i$ is correct. Now given any incorrect model M, $M \cup M_0$ is a correct model. Choose a variable x^* from $M_0 \setminus M$. Clearly, the regression coefficient, say β_1, of x^* is nonzero. By Theorem 3.2,

$$P\left[\mathrm{PLS}_n(M \cup M_0) < \mathrm{PLS}_n(M) \text{ eventually}\right] = 1, \tag{3.55}$$

where $\mathrm{PLS}_n(M)$ denotes the PLS value based on model M. Now given any correct model $M \neq M_0$, there are l, M_i, $1 \le i \le l$, such that $M_l = M$ and for $1 \le i \le l$, $M_i - M_{i-1}$ has only one element. Applying Theorem 3.3 l times, we obtain

$$P\left[\mathrm{PLS}_n(M_0) < \mathrm{PLS}_n(M) \text{ eventually}\right] = 1. \tag{3.56}$$

Combining (3.55) and (3.56), (3.52) is proved. □

REMARK. Theorem 3.4 also holds for the stochastic regression case if (3.51) holds a.s. The only difference between the stochastic and fixed design cases is that (3.30), (3.31) and (3.36) have to be satisfied for the stochastic case. Choose A to be the identity matrix and $\mathbf{z}_n = \mathbf{x}_n$. By the remark immediately following Lemma 3.5, we do not have to verify (3.31). By (3.51), (3.30) holds. Since A is the identity matrix, (3.36) also holds.

EXAMPLE 2 (Unstable autoregressive process). Let us consider the following AR(p) model:

$$y_n = \phi_1 y_{n-1} + \cdots + \phi_p y_{n-p} + \delta_n, \tag{3.57}$$

where $\{\delta_n\}$ satisfies (3.15) and the characteristic polynomial

$$\phi(z) = z^p - \phi_1 z^{p-1} - \cdots - \phi_p \tag{3.58}$$

has all roots either on or inside the unit circle. [This is the unstable case. For statistical properties of unstable AR(p) processes, see Chan and Wei (1988)

and the references therein.] Assume that p is known and $p_0 = \max\{j: \phi_j \neq 0, 1 \leq j \leq p\}$. For each j, let $\mathrm{PLS}_n(j)$ be the PLS value based on fitting an AR(j) model. Then the predictive least squares principle selects $\hat{p}_n$ which satisfies $\mathrm{PLS}_n(\hat{p}_n) = \inf\{\mathrm{PLS}_n(j): 0 \leq j \leq p\}$.

THEOREM 3.5. *Assume that $\{y_n\}$ is an unstable autoregressive process that satisfies* (3.57) *and $y_0, \ldots, y_{-p+1}$ are $\mathcal{F}_0$-measurable. Then*

$$P[\hat{p}_n = p_0 \text{ eventually}] = 1. \tag{3.59}$$

PROOF. For $j \geq p_0$, let $\mathbf{x}_n' = (y_{n-j}, \ldots, y_{n-1})$ and $V_n^{-1} = \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j'$. Since, by the definition of p_0, AR(j) is a correct model, it is known [Lai and Wei (1983)] that

$$\liminf_{n\to\infty} \lambda_*(V_n^{-1})/n > 0 \quad \text{a.s.}, \tag{3.60}$$

$$\limsup_{n\to\infty} \lambda^*(V_n^{-1})/n^k < \infty \quad \text{a.s. for some } k > 0, \tag{3.61}$$

$$\lim_{n\to\infty} \mathbf{x}_n' V_n \mathbf{x}_n = 0 \quad \text{a.s.} \tag{3.62}$$

Now let us show that if $l < p_0$, then

$$P[\mathrm{PLS}_n(l) > \mathrm{PLS}_n(p_0) \text{ eventually}] = 1. \tag{3.63}$$

Since AR(p_0) is a correct model, (3.1) is satisfied. The assumed condition (3.15) implies (3.2). By (3.60) and (3.61), conditions (3.10) and (3.11) also hold. Therefore, in view of the fact that $\phi_{p_0} \neq 0$, (3.63) is a consequence of Theorem 3.2. Next let us show that if $j > p_0$, then

$$P[\mathrm{PLS}_n(j) > \mathrm{PLS}_n(p_0) \text{ eventually}] = 1. \tag{3.64}$$

For this, it is sufficient to prove that if $j > p_0$, then

$$P[\mathrm{PLS}_n(j) > \mathrm{PLS}_n(j-1) \text{ eventually}] = 1. \tag{3.65}$$

Note that by (3.57),

$$y_n = \phi_1 y_{n-1} + \cdots + \phi_j y_{n-j} + \delta_n,$$

where $\phi_j = 0$, since $j > p_0$. Thus condition (3.1) is satisfied. By (3.15) and (3.60)–(3.62), conditions (3.16), (3.22) and (3.42) are also satisfied. To apply Theorem 3.3, it remains to show (3.30), (3.31) and (3.36). Define

$$\varphi(z) = z^j - \phi_1 z^{j-1} - \cdots - \phi_j. \tag{3.66}$$

Since $\phi(z) = z^{p-j}\varphi(z)$ and $\phi(z)$ has all roots inside or on the unit circle, so does φ. Therefore,

$$\varphi(z) = \psi(z)\theta(z), \tag{3.67}$$

where

$$\begin{aligned}&\psi(z) = z^s - \psi_1 z^{s-1} - \cdots - \psi_s \text{ has all roots inside the unit circle,}\\ &\theta(z) = z^t - \theta_1 z^{t-1} - \cdots - \theta_t \text{ has all roots on the unit circle,}\end{aligned} \tag{3.68}$$

and $s + t = j$. Let

$$\begin{aligned}&u_n = y_n - \theta_1 y_{n-1} - \cdots - \theta_t y_{n-t}, \qquad \mathbf{z}_n = (u_{n-s}, \ldots, u_{n-1})',\\ &v_n = y_n - \psi_1 y_{n-1} - \cdots - \psi_s y_{n-s}, \qquad \mathbf{w}_n = (v_{n-t}, \ldots, v_{n-1})'\end{aligned} \tag{3.69}$$

and define the $s \times j$, $r \times j$ and $j \times j$ matrices A_1, A_2 and A by

$$A_1 = \begin{pmatrix} -\theta_t & -\theta_{t-1} & & \cdots & -\theta_1 & 1 & 0 & \cdots & & 0 \\ 0 & -\theta_t & -\theta_{t-1} & \cdots & & -\theta_1 & 1 & 0 & \cdots & \vdots \\ \vdots & 0 & & \cdots & & & & \cdots & & \vdots \\ 0 & & & \cdots & -\theta_t & \theta_{t-1} & & \cdots & -\theta_1 & 1 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} -\psi_s & -\psi_{s-1} & & \cdots & -\psi_1 & 1 & 0 & \cdots & & 0 \\ 0 & -\psi_s & -\psi_{s-1} & \cdots & & -\psi_1 & 1 & 0 & \cdots & \vdots \\ \vdots & 0 & & \cdots & & & & \cdots & & \vdots \\ 0 & & & \cdots & -\psi_s & -\psi_{s-1} & & \cdots & -\psi_1 & 1 \end{pmatrix},$$

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}.$$

Since $\phi_j = 0$, $\psi_s = 0$ and (3.36) is satisfied. Note that $A\mathbf{x}_n = (\mathbf{z}_n', \mathbf{w}_n')'$. It is known [Lai and Wei (1985)] that for some positive-definite matrix Γ,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i' = \Gamma \quad \text{a.s.} \tag{3.70}$$

Hence (3.31) is satisfied. But (3.30) is a corollary of Theorem A.1. Therefore, Theorem 3.3 is applicable and (3.65) is proved. Combining (3.63) and (3.64), we obtain (3.59). □

As a side result of our analysis, we are also able to obtain the strong consistency for BIC. Recall that

$$\mathrm{BIC}(k) = \log \hat{\sigma}_n^2(k) + kn^{-1} \log n,$$

where $n\hat{\sigma}_n^2(k)$ is the residual sum of squares based on an AR(k) fitting.

THEOREM 3.6. *Assume that $\{y_n\}$ is an unstable autoregressive process that satisfies* (3.57) *and $y_0, \ldots, y_{-p+1}$ are $\mathcal{F}_0$-measurable. Let $\hat{k}_n$ be the order that minimizes BIC over $\{0, \ldots, p\}$. Then*

$$P\left[\hat{k}_n = p_0 \text{ eventually}\right] = 1. \tag{3.71}$$

PROOF. Note that for $j \geq p_0$, by (3.61),

$$\lim_{n\to\infty} \log \lambda^*\left(V_n^{-1}\right)/n = 0 \quad \text{a.s.} \tag{3.72}$$

Therefore, by Lemma 3 of Lai and Wei (1982a),

$$\lim_{n\to\infty} \hat{\sigma}_n^2(j) = \sigma^2 \quad \text{a.s.} \tag{3.73}$$

Now if $k < p_0$, then

$$\begin{aligned} &\log \hat{\sigma}_n^2(k) - \log \hat{\sigma}_n^2(p_0) \\ &\quad \geq \log \hat{\sigma}_n^2(p_0 - 1) - \log \hat{\sigma}_n^2(p_0) \\ &\quad = \log\left\{1 + \left[\hat{\sigma}_n^2(p_0)\right]^{-1}\left[\hat{\sigma}_n^2(p_0 - 1) - \hat{\sigma}_n^2(p_0)\right]\right\}. \end{aligned} \tag{3.74}$$

By Theorem 3.1,

$$n\left[\hat{\sigma}_n^2(p_0 - 1) - \hat{\sigma}_n^2(p_0)\right] \geq \left[\sum_{i=1}^{n} \delta_i^2(n) - \sum_{i=1}^{n} \delta_i^2\right] \sim s_n^2\beta_1^2 \quad \text{a.s.} \tag{3.75}$$

By (3.9), (3.60) and (3.75),

$$\liminf_{n\to\infty}\left[\hat{\sigma}^2(p_0 - 1) - \hat{\sigma}_n^2(p_0)\right] > 0 \quad \text{a.s.}$$

This, (3.73) and (3.74) in turn imply that

$$\begin{aligned} &\liminf_{n\to\infty}\left[\mathrm{BIC}(k) - \mathrm{BIC}(p_0)\right] \\ &\quad \geq \liminf_{n\to\infty}\left[\log \hat{\sigma}_n^2(k) - \log \hat{\sigma}_n^2(p_0)\right] > 0 \quad \text{a.s.} \end{aligned} \tag{3.76}$$

Now if $l > p_0$, by (3.37) of Corollary 3.1 and (3.61),

$$n\left[\hat{\sigma}_n^2(l - 1) - \hat{\sigma}_n^2(l)\right] = o(\log n) \quad \text{a.s.}$$

This and (3.73) imply that

$$\begin{aligned} &n\left[\log \hat{\sigma}_n^2(l - 1) - \log \hat{\sigma}_n^2(l)\right] \\ &\quad = n \log\left\{1 + \left[\hat{\sigma}_n^2(l - 1)\right]^{-1}\left[\hat{\sigma}_n^2(l - 1) - \hat{\sigma}_n^2(l)\right]\right\} \\ &\quad \leq n\left[\hat{\sigma}_n^2(l - 1)\right]^{-1}\left[\hat{\sigma}_n^2(l - 1) - \hat{\sigma}_n^2(l)\right] = o(\log n) \quad \text{a.s.} \end{aligned}$$

Consequently,

$$\lim_{n\to\infty} n\left[\mathrm{BIC}(l) - \mathrm{BIC}(l - 1)\right]/(\log n) = 1 \quad \text{a.s.},$$

and

$$\lim_{n\to\infty} n\left[\mathrm{BIC}(l) - \mathrm{BIC}(p_0)\right]/(\log n) = l - p_0 \quad \text{a.s.} \tag{3.77}$$

Combining (3.76) and (3.77), (3.71) is proved. □

REMARK. The criterion PLS is different from BIC. For example, if $p = 1$ and $\phi_1 = 1$ in (3.57), then by (2.10), (3.17) and (3.5) of Wei (1987),

$$\text{PLS} \sim n\hat{\sigma}_n^2(1) + \sigma^2 \log\left(\sum_{i=1}^{n-1} y_i^2\right)$$

$$\sim n\hat{\sigma}_n^2(1) + 2\sigma^2 \log n.$$

Consequently,

$$\log[\text{PLS}/n] \sim \log \hat{\sigma}_n^2 + 2 \log n = \text{BIC} + \log n.$$

In general, the penalty term depends on the number of roots of (3.58) which are inside and on the unit circle, see (3.11) of Wei (1987).

4. Incorrect models. In this section we study the asymptotic decomposition of the PLS when the model is incorrect. For this, we introduce $\hat{\sigma}_n^2 = (1/n)\sum_{i=1}^n \hat{\varepsilon}_i^2(n)$. In the following discussion, regression and time series models will be treated separately.

4.1. *Regression model.* Consider the regression model

$$y_i = f((X_i)) + \alpha_i, \tag{4.1.1}$$

where $X_i = (x_{i1}, x_{i2}, \dots) \in l^2$, f is a measurable function from l^2 into R and α_i are i.i.d. random variables with $E(\alpha_i) = 0$ and $\text{var}(\alpha_i) = \sigma^2 > 0$. One much discussed regression function [see Shibata (1981)] is of the form

$$f(X_i) = \sum_{j=1}^{\infty} \theta_j x_{ij}, \quad \text{where } (\theta_1, \theta_2, \dots) \in l^2. \tag{4.1.2}$$

Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$. The following theorem gives the PLS an asymptotic decomposition when p variables $\mathbf{x}_i$ are selected.

THEOREM 4.1.1. *Assume that* (4.1.1) *holds. If*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' = \Gamma \quad \textit{for some nonsingular } \Gamma, \tag{4.1.3}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i f(X_i) = \boldsymbol{\gamma}, \tag{4.1.4}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' [f(X_i) - \boldsymbol{\beta}'\mathbf{x}_i]^2 = \tilde{G}, \quad \textit{where } \boldsymbol{\beta} = \Gamma^{-1}\boldsymbol{\gamma}, \tag{4.1.5}$$

$$\sup_n \|\mathbf{x}_n\| < \infty, \tag{4.1.6}$$

then

$$(4.1.7)\qquad \sum_{i=m+1}^{n} e_i^2 = n\hat{\sigma}_n^2 + (\log n)\left[p\sigma^2 + \operatorname{tr}(\Gamma^{-1}\tilde{G})\right](1 + o(1)) \quad a.s.$$

PROOF. Let $\varepsilon_i = y_i - \boldsymbol{\beta}'\mathbf{x}_i$. Then $\varepsilon_i = (f(X_i) - \boldsymbol{\beta}'\mathbf{x}_i) + \alpha_i$. We will apply Theorem 2.3 with $h_i = f(X_i) - \boldsymbol{\beta}'\mathbf{x}_i$. Observe that (4.1.3) is exactly the same as (2.14). By (4.1.5), to show (2.21), we only have to show

$$(4.1.8)\qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\alpha_i^2 = \sigma^2\Gamma.$$

If $|\alpha_i| \le K$ a.s., then by Chow's (1965) result and (4.1.6),

$$\sum_{i=1}^{n} x_{ij}x_{il}(\alpha_i^2 - \sigma^2) = o\left(\sum_{i=1}^{n} \|\mathbf{x}_i\|^4 K^4\right) = o(n) \quad \text{a.s.}$$

This and (4.1.3) imply (4.1.8). For the general case, let $\tilde{\alpha}_i = \alpha_i I_{[|\alpha_i| \le K]}$. Then by (4.1.8) and the strong law of large numbers,

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n}\left|\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'(\alpha_i^2 - \tilde{\alpha}_i^2)\right| &\le \left(\sup_n \|x_n\|^2\right) \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \alpha_i^2 I_{[|\alpha_i| > K]} \\ &= \left(\sup_n \|x_n\|^2\right) E\alpha_1^2 I_{[|\alpha_1| > K]}.\end{aligned}$$

Applying the result for the bounded case, we obtain (4.1.8) by letting $K \to \infty$. Now let us prove (2.22). By Chow's theorem again,

$$\sum_{i=m+1}^{n} \mathbf{x}_i'V_i\mathbf{x}_i h_i\alpha_i = O\left(\sum_{i=m+1}^{n} (\mathbf{x}_i'V_i\mathbf{x}_i h_i)^2 E\alpha_i^2\right) = O\left(\sum_{i=m+1}^{n} (\mathbf{x}_i'V_i\mathbf{x}_i)^2 h_i^2\right).$$

Finally, let us verify (2.24). It is known [Lai, Robbins and Wei (1979)] that

$$\left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i\alpha_i = o(1) \quad \text{a.s.}$$

This in turn implies that

$$\begin{aligned}\mathbf{b}_n - \boldsymbol{\beta} &= \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i - \boldsymbol{\beta} \\ &= \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i f(X_i) - \boldsymbol{\beta} + \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i\alpha_i \\ &\to \Gamma^{-1}\boldsymbol{\gamma} - \boldsymbol{\beta} = \mathbf{0} \quad \text{a.s.}\end{aligned}$$

Hence, by (4.1.6),

$$\|\mathbf{x}_n'(\mathbf{b}_{n-1} - \boldsymbol{\beta})\| \le \left(\sup_n \|\mathbf{x}_n\|\right)\|\mathbf{b}_{n-1} - \boldsymbol{\beta}\| = o(1) \quad \text{a.s.} \qquad \square$$

REMARK. Theorem 4.1 also holds for arbitrary p variables of $(x_{i1}, x_{i2}, \dots)$.

EXAMPLE. Polynomial regression. Let

$$f(x) = \sum_{j=1}^{\infty} \theta_j x^{j-1},$$

where $(\theta_1, \theta_2, \dots) \in l^2$ and $0 \le x < 1$. This is a special case of (4.1.2) with $x_{ij} = (x_i)^{j-1}$. Assume that $\{x_i\}$ is chosen so that there is a distribution F and for any continuous point x of F,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I_{[x_i \le x]} = F(x).$$

Assume that f is bounded. Let $\mathbf{x} = (x^{i_1}, \dots, x^{i_p})$ be the selected variables. Then $\Gamma = \int \mathbf{x}\mathbf{x}'\, dF$, $\boldsymbol{\gamma} = \int \mathbf{x} f(x)\, dx$, $\tilde{G} = \int \mathbf{x}\mathbf{x}'(f(x) - \boldsymbol{\beta}'\mathbf{x})^2\, dF$. Clearly, (4.16) is satisfied with $\sup_n \|\mathbf{x}_n\| \le \sqrt{p}$. The only assumption we have to impose on F is that Γ is nonsingular.

REMARK. To check whether PLS is asymptotically equivalent to BIC, by (1.7), it is equivalent to check whether

$$\lim_{n\to\infty} \frac{p\sigma^2 + \operatorname{tr}(\Gamma^{-1}\tilde{G})}{\hat{\sigma}_n^2} = p \quad \text{a.s.} \tag{4.1.9}$$

Assume that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f^2(X_i) = c^2. \tag{4.1.10}$$

Then by (4.1.3), (4.1.4) and (4.1.10),

$$\begin{aligned} \lim_{n\to\infty} \hat{\sigma}_n^2 &= \sigma^2 + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left[f(X_i) - \boldsymbol{\beta}'\mathbf{x}_i \right]^2 \\ &= \sigma^2 + c^2 - 2\boldsymbol{\beta}'\boldsymbol{\gamma} + \boldsymbol{\beta}'\Gamma\boldsymbol{\beta}. \end{aligned}$$

Therefore, (4.1.9) is equivalent to

$$\operatorname{tr}(\Gamma^{-1}\tilde{G}) = p\left[c^2 - 2\boldsymbol{\beta}'\boldsymbol{\gamma} + \boldsymbol{\beta}'\Gamma\boldsymbol{\beta}\right]. \tag{4.1.11}$$

In the polynomial regression case, let F be the uniform distribution over $[0, 1)$ and $f(x) \equiv 1$. Assume that the variable $\mathbf{x} = (x)$ is the selected variable. Then $\Gamma = 1/3$, $\boldsymbol{\gamma} = 1/2$, $\boldsymbol{\beta} = 3/2$,

$$\tilde{G} = \int_0^1 x^2(1 - (3/2)x)^2\, dx = 1/30, \qquad \operatorname{tr}(\Gamma^{-1}\tilde{G}) = 1/10$$

and

$$p(c^2 - 2\boldsymbol{\beta}'\boldsymbol{\gamma} + \boldsymbol{\beta}'\Gamma\boldsymbol{\beta}) = \int_0^1 (1 - (3/2)x)^2\, dx = 1/4 \ne 1/10.$$

Clearly, PLS is not asymptotically equivalent to BIC.

4.2. *Time series.* In this section we assume that

$$(y_n, \mathbf{x}_n) \text{ is ergodic and } E(|y_1| + \|\mathbf{x}_1\|)^4 < \infty, \tag{4.2.1}$$

$$E(\mathbf{x}_1\mathbf{x}_1') = \Gamma \text{ is nonsingular.} \tag{4.2.2}$$

Let $\boldsymbol{\gamma} = E(\mathbf{x}_1 y_1)$, $\boldsymbol{\beta} = \Gamma^{-1}\boldsymbol{\gamma}$, $\varepsilon_i = y_i - \boldsymbol{\beta}'\mathbf{x}_i$ and $G = E(\mathbf{x}_1\mathbf{x}_1'\varepsilon_1^2)$.

THEOREM 4.2.1. *Assume that* (4.2.1) *and* (4.2.2) *hold. Then*

$$\sum_{i=m+1}^{n} e_i^2 = n\hat{\sigma}_n^2 + (\log n)\operatorname{tr}(\Gamma^{-1}G)(1 + o(1)) \quad a.s., \tag{4.2.3}$$

if one of the following conditions is satisfied:

$$\sup_n \|\mathbf{x}_n\| < \infty \quad a.s.; \tag{4.2.4}$$

$$\begin{aligned}&(y_n, \mathbf{x}_n) \text{ is strong mixing with mixing rate } \alpha(n) \text{ and}\\ &\text{there exist } r > 2,\ \delta > 0, \text{ such that } \textstyle\sum_{i=1}^{\infty} i^{r/2-1}[\alpha(i)]^{\delta/(r+\delta)}\\ &< \infty \text{ and } E(|y_1| + \|\mathbf{x}_1\|)^{2(r+\delta)} < \infty;\end{aligned} \tag{4.2.5}$$

$$\begin{aligned}&\text{there exist i.i.d. mean zero random variables } \{\delta_n\},\\ &\text{sequences } \{a_n\},\ \{b_n(j)\} \text{ and } r > 4, \text{ such that } E|\delta_1|^r < \infty,\\ &\textstyle\sum_{-\infty}^{\infty}(|a_n| + |b_n(j)|) < \infty,\quad y_n = \sum_{-\infty}^{\infty} a_{n-i}\delta_i \quad \text{and} \quad x_{nj} =\\ &\textstyle\sum_{-\infty}^{\infty} b_{n-i}(j)\delta_i.\end{aligned} \tag{4.2.6}$$

PROOF. We will prove that the conditions of Theorem 2.3 hold with $h_i = \varepsilon_i$, $\alpha_i = 0$ (hence $\sigma^2 = 0$) and $\tilde{G} = G$. Condition (2.14) and the first part of (2.21) follow from the ergodic theorem. The second part of (2.21) and (2.22) are immediate consequences of the fact that $\alpha_i = 0$. It only remains to show (2.24) or

$$\lim_{n\to\infty} \mathbf{x}_n'(\mathbf{b}_{n-1} - \boldsymbol{\beta}) = 0 \quad \text{a.s.} \tag{4.2.7}$$

Now by the ergodic theorem

$$\lim_{n\to\infty} \mathbf{b}_n = \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i = \left[E(\mathbf{x}_1\mathbf{x}_1')\right]^{-1} E(\mathbf{x}_1 y_1) = \boldsymbol{\beta} \quad \text{a.s.}$$

Therefore, under (4.2.4),

$$\|\mathbf{x}_n'(\mathbf{b}_{n-1} - \boldsymbol{\beta})\| \le \left(\sup_n \|\mathbf{x}_n\|\right)\|\mathbf{b}_{n-1} - \boldsymbol{\beta}\| = o(1) \quad \text{a.s.}$$

To show that (4.2.7) holds under (4.2.5) or (4.2.6), observe first that $E\|\mathbf{x}_1\|^4 < \infty$ implies $\|\mathbf{x}_n\| = o(n^{1/4})$. Hence it is sufficient to show that

$$\|\mathbf{b}_n - \boldsymbol{\beta}\| = O(n^{-1/4}) \quad \text{a.s.} \tag{4.2.8}$$

Note that

$$\mathbf{b}_n - \boldsymbol{\beta} = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i (y_i - \boldsymbol{\beta}' \mathbf{x}_i).$$

By the ergodic theorem, proving (4.2.8) is equivalent to proving

$$\sum_{i=1}^{n} \mathbf{x}_i (y_i - \boldsymbol{\beta}' \mathbf{x}_i) = \sum_{i=1}^{n} \mathbf{x}_i \varepsilon_i = O(n^{3/4}) \quad \text{a.s.} \tag{4.2.9}$$

Now fix j and let $Z_i = x_{ij}\varepsilon_i$. Since

$$E(\mathbf{x}_1 \varepsilon_1) = E(\mathbf{x}_1 y_1) - E(\mathbf{x}_1 \mathbf{x}_1')\boldsymbol{\beta} = \boldsymbol{\gamma} - \Gamma(\Gamma^{-1}\boldsymbol{\gamma}) = \mathbf{0}, \tag{4.2.10}$$

$E(Z_i) = 0$. Under (4.2.5), we have $E|Z_1|^{r+\delta} < \infty$ and the mixing rate $\tilde{\alpha}(n)$ of $\{Z_n\}$ satisfies

$$\sum_{i=1}^{\infty} i^{r/2-1} [\tilde{\alpha}(i)]^{\delta/(r+\delta)} < \infty,$$

since obviously $\tilde{\alpha}(n) \leq \alpha(n)$. Consequently, by Corollary 2 of Yokoyama (1980), page 47,

$$\sum_{i=1}^{n} Z_i = o(n^{1/2} \log n) \quad \text{a.s.}, \tag{4.2.11}$$

and (4.2.9) holds. Now let us assume (4.2.6). By Lemma A.1 and the fact that $E(Z_i) = 0$,

$$E\left| \sum_{i=1}^{n} Z_i \right|^{r/2} \leq K n^{r/4} \quad \text{for some } K > 0.$$

Since $r/4 > 1$, by Corollary 1 of Lai and Wei (1984), (4.2.11) holds and so does (4.2.9). □

EXAMPLE. Let y_n be an AR(∞) process defined by

$$\sum_{j=0}^{\infty} \phi_j y_{n-j} = \delta_n,$$

where $\phi_0 = 1$, $\sum_{j=0}^{\infty} |\phi_j| < \infty$, $\phi(z) = \sum_{j=0}^{\infty} \phi_j z^j \neq 0$ for $|z| \leq 1$ and δ_n are i.i.d. random variables with $E(\delta_1) = 0$, $E(\delta_1^2) = \sigma^2$ and $E|\delta_1|^r < \infty$ for some $r > 4$. We fit the model by an AR(p) model. Hence $\mathbf{x}_n = (y_{n-1}, \ldots, y_{n-p})$. By a theorem due to Wiener and Lévy [Zygmund (1959), page 245], there is a sequence $\{a_n\}$ such that $\sum |a_n| < \infty$ and $y_n = \sum_{i=0}^{\infty} a_i \delta_{n-i}$. Therefore, (4.2.6) holds.

Let $\gamma(i)$ be the covariance function of y_n. We then have that $\Gamma = (\gamma(i-j))_{i,j=1,\ldots,p}$, $\boldsymbol{\gamma} = (\gamma(1), \ldots, \gamma(p))'$, $\boldsymbol{\beta} = \Gamma^{-1}\boldsymbol{\gamma}$. Furthermore, by the inde-

pendence of δ_n and $\{y_j\colon j < n\}$,

$$G = E\mathbf{x}_1\mathbf{x}_1'\left(y_2 - \sum_{j=1}^{p}\beta_j y_{2-j}\right)^2$$

$$= E\mathbf{x}_1\mathbf{x}_1'\left(-\sum_{j=1}^{\infty}\phi_j y_{2-j} - \sum_{j=1}^{p}\beta_j y_{2-j}\right)^2 + \sigma^2 E(\mathbf{x}_1\mathbf{x}_1')$$

$$= G_1 + \sigma^2\Gamma \quad \text{(say)}.$$

Therefore, $\operatorname{tr}(\Gamma^{-1}G) = p\sigma^2 + \operatorname{tr}(\Gamma^{-1}G_1)$.

REMARK. Under slightly stronger assumptions on ϕ_j and δ_j, Kavalieris (1989), Theorem 1, claimed that in an AR(∞) model

$$\sum_{i=m+1}^{n} e_i^2 = n\hat{\sigma}_n^2 + (\log n)\left(p\sigma_p^2 + C_p\right)(1 + o(1)) \quad \text{a.s.}, \tag{4.2.12}$$

where

$$\sigma_p^2 = E\left(y_2 - \sum_{j=1}^{p}\beta_j y_{2-j}\right)^2 = \sigma^2 + E\left(\sum_{j=1}^{\infty}\phi_1 y_{2-j} - \sum_{j=1}^{p}\beta_j y_{2-j}\right)^2$$

and $0 \le C_p \le C(\sigma_p^2 - \sigma^2)$ for some C.

This does not coincide with our result. Let us check a simple example where $y_n = \delta_n + \rho\delta_{n-1}$ with $|\rho| < 1$. Consider fitting this process by an AR(1) model. We obtain

$$\Gamma = (1 + \rho^2)\sigma^2, \qquad \gamma = \rho\sigma^2, \qquad \beta = \rho/(1 + \rho^2).$$

Hence

$$\sigma_p^2 = E(y_2 - \beta y_1)^2 = \sigma^2 + (\rho - \beta)^2\sigma^2 + \beta^2\rho^2\sigma^2 = \sigma^2 + \sigma^2\rho^4/(1 + \rho^2).$$

But

$$G_1 = E\left[y_1^2(y_2 - \beta y_1 - \delta_2)^2\right]$$

$$= \left[2\rho^6 E(\delta_1^4) + (\rho^8 - 4\rho^6 + \rho^4)\sigma^4\right]/(1 + \rho^2)^2$$

$$= 2\rho^6\left[E(\delta_1^4) - 3\sigma^4\right]/(1 + \rho^2) + \rho^4\sigma^4.$$

Therefore,

$$\begin{aligned}\operatorname{tr}(\Gamma^{-1}G) &= \sigma^2 + \operatorname{tr}(\Gamma^{-1}G_1)\\ &= \sigma_p^2 + \left\{2\rho^6/\left[(1 + \rho^2)\sigma^2\right]\right\}\left[E(\delta_1^4) - 3\sigma^4\right].\end{aligned} \tag{4.2.13}$$

Clearly, when $E(\delta_1^4) < 3\sigma^4$, $\operatorname{tr}(\Gamma^{-1}G) < \sigma_p^2 + C_p$. This is the case when $P[\delta_1 = 1] = P[\delta_1 = -1] = 1/2$.

In the above discussion, we find that in general $\operatorname{tr}(\Gamma^{-1}G)$ may not be equivalent to σ_p^2. Since $\lim_{n\to\infty}\hat{\sigma}_n^2 = \sigma^2$ a.s. [see (4.2.17)], this in turn implies

that PLS asymptotically may not be equal to BIC. However, by (4.2.13), if $E(\delta_1^{y4}) = 3\sigma^4$, then $\operatorname{tr}(\Gamma^{-1}G) = \sigma_p^2$ and PLS is expected to be equal to BIC. This is not an accident. We have the following result.

THEOREM 4.2.2. *Assume that* (4.2.1), (4.2.2) *and* (4.2.6) *hold. If*

$$E(\delta_1^4) = 3\sigma^4, \tag{4.2.14}$$

then

$$\operatorname{tr}(\Gamma^{-1}G) = pE(\varepsilon_1^2) = pE(y_1 - \boldsymbol{\beta}'\mathbf{x}_1)^2 \tag{4.2.15}$$

and

$$\log\left(\frac{1}{n}\sum_{i=m+1}^{n} e_i^2\right) = \log\hat{\sigma}_n^2 + p\left(\frac{\log n}{n}\right)(1 + o(1)) \quad a.s. \tag{4.2.16}$$

COROLLARY 4.2.1. *Assume that* (4.2.1), (4.2.2) *and* (4.2.6) *hold. If* δ_i *is Gaussian, then* (4.2.15) *and* (4.2.16) *hold.*

The corollary follows immediately from Theorem 4.2.2 since a normal random variable satisfies (4.2.14).

PROOF OF THEOREM 4.2.2. Since the calculation of Γ and G only involves σ^2 and $E(\delta_1^4)$, without loss of generality, we can assume that δ_n is Gaussian. By (4.2.10) and the Gaussianity of $(\mathbf{x}_1, \varepsilon_1)$, $\mathbf{x}_1$ is independent of ε_1. Therefore,

$$G = E(\mathbf{x}_1\mathbf{x}_1'\varepsilon_1^2) = E(\mathbf{x}_1\mathbf{x}_1')E(\varepsilon_1^2) = \Gamma E(\varepsilon_1^2)$$

and

$$\operatorname{tr}(\Gamma^{-1}G) = \operatorname{tr}(\Gamma^{-1}\Gamma)E(\varepsilon_1^2) = pE(\varepsilon_1^2).$$

Therefore, (4.2.15) is proved. By the ergodic theorem and (4.2.8),

$$\begin{aligned}
\hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{b}_n'\mathbf{x}_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2 - \frac{1}{n}(\mathbf{b}_n - \boldsymbol{\beta})'\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{b}_n - \boldsymbol{\beta}) \qquad (4.2.17)\\
&= E(\varepsilon_1^2)(1 + o(1)) + O(\|\mathbf{b}_n - \boldsymbol{\beta}\|^2) \\
&= E(\varepsilon_1^2) + o(1) \quad \text{a.s.}
\end{aligned}$$

By (4.2.3), (4.2.15) and (4.2.17),

$$\begin{aligned}
\log\left(\frac{1}{n}\sum_{i=m+1}^{n} e_i^2\right) &= \log\hat{\sigma}_n^2 + \log\left[1 + p\frac{E(\varepsilon_1^2)}{\hat{\sigma}_n^2}\frac{\log n}{n}(1 + o(1))\right] \\
&= \log\hat{\sigma}_n^2 + \log\left[1 + p\frac{\log n}{n}(1 + o(1))\right] \\
&= \log\hat{\sigma}_n^2 + p\frac{\log n}{n}(1 + o(1)) \quad \text{a.s.}
\end{aligned}$$

This completes our proof. □

REMARK. If (4.2.1), (4.2.2) and (4.2.5) are satisfied and $(y_n, \mathbf{x}_n)$ is Gaussian, then by the same proof, (4.2.15) and (4.2.16) also hold.

5. A Fisher information criterion. In this section we propose a new model selection criterion that is based on the Fisher information. We first give this new criterion a statistical interpretation and discuss its relationship with the predictive least squares principles. We then use a simulation study to demonstrate its advantage.

5.1. *FIC.* The criterion we propose is to select the model that minimizes

$$\text{FIC}(M) = n\hat{\sigma}_n^2 + \tilde{\sigma}_n^2 \log\det\left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\right), \tag{5.1.1}$$

where M is the model with design vector $\mathbf{x}_i$ and $\hat{\sigma}_n^2$ and $\tilde{\sigma}_n^2$ are variance estimators based on the model M and the full model, respectively.

When $y_i = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i$, where ε_i are i.i.d. $N(0, \sigma^2)$ and $\mathbf{x}_i$ is $\sigma(\varepsilon_1, \ldots, \varepsilon_{i-1})$-measurable, the conditional Fisher information matrix for $\boldsymbol{\beta}$ is $\sigma^{-2}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'$. The quantity $\det(\sigma^{-2}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i')$ can be interpreted as the amount of information about $\boldsymbol{\beta}$. Our criterion replaces the conventional penalty term, which is proportional to the *topological dimension* of the selected model, by a term that is proportional to the logarithm of the *statistical information* contained in M. The redundant information by introducing a spurious variable is used to represent its penalty. When this variable, say x_1, is uncorrelated to other variables [so that in (3.19) $s_n^2 \sim \sum_{i=1}^{n} x_{i1}^2$], the penalty term is larger if the magnitude of this variable is bigger. This is a desirable property since the prediction error for the variable with larger magnitude is expected to be bigger. The conventional criteria do not possess such a feature. Furthermore, when an l-variate AR(p) model is fitted, one may have conceptual difficulty deciding whether p or lp should be used in the conventional criteria. Our criterion does not have this ambiguity.

The relationship between FIC and PLS is very close. In the stochastic regression model, using (2.6) and Chow's (1965) result, one can see that

$$\begin{aligned}\text{PLS} \sim{}& n\hat{\sigma}_n^2 + \sigma^2 \log\det\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)\\ &+ \sum_{i=m+1}^{n}\mathbf{x}_i'V\mathbf{x}_i\left[\mathbf{x}_i'(\mathbf{b}_{n-1} - \boldsymbol{\beta})\right]^2 \quad \text{a.s.}\end{aligned} \tag{5.1.2}$$

When the model is correct, under the assumptions given in Lemmas 3.1 and 3.2, the last term can be dropped and we have

$$\text{PLS} \sim n\hat{\sigma}_n^2 + \sigma^2 \log\det\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right). \tag{5.1.3}$$

Replacing σ^2 in (5.1.3) by an estimator, say $\tilde{\sigma}_n^2$, we then obtain FIC. When the model is incorrect, in comparison with FIC, PLS has an extra penalty term.

However, this gain seems to be minimal. Because in this case, $n\hat{\sigma}_n^2$ is the dominant term. Furthermore, PLS has the following unpleasant features in practice.

(1) *It is computer intensive.* Although there are some algorithms [Friedlander (1982) and Wax (1988)] to reduce the size of the computation operations, due to its recursive nature, it is still very computer intensive. For a discussion, see Hannan, Mcdougall and Poskit (1989).

(2) *It tends to select the model with fewer variables when the sample size is small.* This is observed by Rissanen [(1986c), page 60] and a philosophical explanation is also given there. According to the simulation study below, we observe that at earlier stages, the model with more variables tends to have larger prediction errors. When the sample size is small, these errors tend to dominate the PLS. Consequently, it rejects the complicated models more frequently than it should. Instead of computing PLS starting from $m + 1$ when the first prediction error is well defined, one may start at a later stage, say $\tilde{m}$. Presumably, this would improve the performance of the prediction for the model with more variables and resolve the problem mentioned above. However, when the sample size is small, it is not clear how to choose such a cut point $\tilde{m}$.

(3) *It is a criterion that depends on the particular order of data.* Although using PLS as an on-line model selection procedure is quite natural since the data come in sequentially, conceptually it may not be that sound for the off-line case. For example, if $\{y_i\}$ is a stationary Gaussian process, then $(y_1, y_2, \ldots, y_n)$ and $(y_n, y_{n-1}, \ldots, y_1)$ have the same distribution; that is, the process is reversible. In this situation, when one observes $\{y_1, \ldots, y_n\}$, should he or she use the time-ordered data or its reverse? Furthermore, as recognized by Rissanen (1986a, b), when the data are modeled as being independent, a permutation-invariant criterion is required. Since using all permuted sequences to compute PLS is a formidable task, Rissanen also suggests a modification [see (2.6) of Rissanen (1986a)] of the PLS. However, the computation of this modification, although not of exponential complexity, is much more involved than the computation of any particular PLS. Furthermore, this modification is a local optimal procedure. It is not clear, at least conceptually, that it would perform as well as the one by all permutations.

Our criterion FIC is permutation invariant, easy to compute and of no initialization problem as stated in (2). It seems that FIC provides a resolution to the above-mentioned problems. Furthermore, FIC also shares strong consistency properties. This is the context of the following theorem.

THEOREM 5.1.1. *Assume that* (3.1), (3.2), (3.10) *and* (3.11) *hold. Then* (3.13) *holds if one replaces PLS by FIC. Furthermore, assume that* (3.1), (3.15), (3.16), *and* (3.22) *hold. Also assume that either* $\mathbf{x}_n$ *are nonrandom vectors or* (3.30), (3.31) *and* (3.36) *hold. If* $\beta_1 = 0$ *and* (3.42) *is replaced by*

$$\lim_{n\to\infty} \log \lambda^*\left(V_n^{-1}\right)/\lambda_*\left(V_n^{-1}\right) = 0 \quad a.s., \tag{5.1.4}$$

and

$$\lim_{n\to\infty} \log \lambda^*\left(V_n^{-1}\right)/n = 0 \quad a.s., \tag{5.1.5}$$

then (3.43) *also holds for FIC.*

Note that (5.1.4) is weaker than (3.42). The reason for introducing (3.42) is to handle the $o(1)$ term appearing in (3.44) and (3.45). The condition (5.1.5) is to ensure that $\tilde{\sigma}_n^2$ in the FIC would converge to σ^2 a.s. [see Lai and Wei (1982a)]. The proof of Theorem 5.1 follows the same arguments as those given in Theorems 3.2 and 3.3. We omit it. Note that FIC (5.1.1) is a special case considered by Pötscher (1989). But in his consistency results, the penalty term is assumed to be of order larger than $\log\det(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i')$. Therefore, his results are not directly applicable here. Also note that the two examples, all subset selection for the regression model and order determination for an AR(p) process, considered in Section 3 both satisfy (5.1.5). Hence FIC also picks up the desired models consistently.

5.2. *Simulation study.* The criterion has been applied to the motor vehicle death data [Draper and Smith (1981), page 191]. The variable x_4, which reports whether there are more males than females, is not included in our study. All chosen criteria, FPE, AIC, BIC, C_p and FIC select the same variables x_1, x_3 and x_5. We also apply these criteria to the cement data [Hald (1952), page 647]. Our criterion FIC, which coincides with BIC and C_p, selects $\{x_1, x_2\}$ while FPE and AIC select $\{x_1, x_2, x_4\}$. It appears that FIC is quite comparable with respect to the other criteria.

To demonstrate the advantage of using FIC, we also conduct a simulation study. The model we consider is

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \delta_i, \qquad i = 1, \ldots, n, \tag{5.2.1}$$

where $x_{i0} = 1$, $x_{i1} = i$, $x_{i2} = \sum_{j=1}^i u_j$, $\{u_i\}$ and $\{\delta_i\}$ are independent sequences of i.i.d. $N(0, 1)$ random variables. In economics [Phillips (1986)], one may like to see whether the data fit a constant model M_0 (i.e., $\beta_2 = \beta_3 = 0$) or a linear

TABLE 1
$\beta_0 = \beta_1 = 1, \beta_2 = 0$ (M_1 *is true*)

Model	FPE	AIC	BIC	C_p	PLS	FIC
M_0	0	0	0	0	0	0
M_1	83	83	86	86	98	93
M_2	0	0	0	0	0	0
M_3	17	17	14	14	2	7

TABLE 2
$\beta_0 = \beta_2 = 1, \beta_1 = 0$ (M_2 *is true*)

Model	FPE	AIC	BIC	C_p	PLS	FIC
M_0	0	0	0	0	0	0
M_1	0	0	0	0	0	0
M_2	86	86	88	87	99	98
M_3	14	14	12	13	1	2

trend model M_1 (i.e., $\beta_1 \neq 0$, $\beta_2 = 0$) or a random walk model M_2 (i.e., $\beta_1 = 0$, $\beta_2 \neq 0$) or a mixture of them, M_3 (i.e., the full model). Note that in this problem, $\Sigma_{i=1}^{n} x_{i0}^2 = n$, $\Sigma_{i=1}^{n} x_{i1}^2 \sim n^3/3$ and $E(\Sigma_{i=1}^{n} x_{i2}^2) = n(n-1)/2$. Each variable has a magnitude of different order. Our criterion FIC (or PLS) is expected to perform better than the other criteria.

Tables 1 to 3 summarize our study. There are three cases. (In regression analysis, one always includes β_0.) In each case, we choose sample size $n = 50$ and run 100 replicants. Each entry in the tables represents the frequency of the model selected by a particular criterion.

As expected, the overall performance of FIC is better than the conventional criteria, since it is sensitive to the magnitude of the selected variable. When the model has fewer parameters (say M_1 and M_2), PLS is better than FIC. However, for the full model, PLS is the worst among all the criteria. When we check into the details, we find that most of the unsuccessful samples for PLS have large initial prediction errors. These errors carry over as the PLS is calculated recursively and become the dominant factor as described in (2) above. Overall, FIC is the best among all the criteria.

5.3. *Final remark.* One may find that FIC is not invariant under the scalar change of $\mathbf{x}_i$. But this can easily be resolved if $\det(\Sigma_{v=1}^{n} \mathbf{x}_i \mathbf{x}_i')$ is replaced by $\det(\hat{\sigma}_n^{-2} \Sigma_{v=1}^{n} \mathbf{x}_i \mathbf{x}_i')$ or $[\det(\Sigma_{v=1}^{n} \mathbf{x}_i \mathbf{x}_i') - \det(\Sigma_{v=1}^{m} \mathbf{x}_i \mathbf{x}_i')]$ for some $m < n$. Theorem 5.1.1. still holds under this modification.

TABLE 3
$\beta_0 = \beta_1 = \beta_2 = 1$ (M_3 *is true*)

Model	FPE	AIC	BIC	C_p	PLS	FIC
M_0	0	0	0	0	0	0
M_1	0	0	0	0	11	0
M_2	0	0	0	0	0	0
M_3	100	100	100	100	89	100

APPENDIX A

Unstable autoregressive process. In this section we are going to show that (3.30) holds for $\mathbf{z}_n$ and $\mathbf{w}_n$ defined by (3.69). By (3.68) and (3.69),

$$u_n = \psi_1 u_{n-1} + \cdots + \psi_s u_{n-s} + \delta_n \tag{A.1}$$

and

$$v_n = \theta_1 v_{n-1} + \cdots + \theta_t v_{n-t} + \delta_n, \tag{A.2}$$

where $\psi(z)$ and $\theta(z)$ satisfy (3.68). Therefore, (3.69) is a corollary of the following theorem.

THEOREM A.1. *Assume that* $\{\delta_n, \mathscr{F}_n\}$ *is a sequence of martingale differences that satisfies* (3.15). *Assume that* u_n *and* v_n *satisfy* (A.1) *and* (A.2). *Let* $\mathbf{u}_n = (u_{n-1}, \ldots, u_{n-s})'$, $\mathbf{v}_n = (v_{n-1}, \ldots, v_{n-t})'$, $P_n = \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_1'$ *and* $H_n = \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i'$. *If* $\mathbf{u}_0$ *and* $\mathbf{v}_0$ *are* $\mathscr{F}_0$*-measurable and* $\psi(z)$ *and* $\theta(z)$ *satisfy* (3.68), *then*

$$\lim_{n\to\infty} P_n^{-1/2}\left(\sum_{i=1}^n \mathbf{u}_i \mathbf{v}_i'\right) H_n^{-1/2} = \mathbf{0} \quad a.s. \tag{A.3}$$

Before we prove Theorem A.1, we need a few lemmas which are of independent interest themselves.

LEMMA A.1. *Let* $\{\mathscr{F}_n\}$ *be a sequence of increasing* σ*-fields. Assume that* $\{\mathbf{y}_n\}$, $\{\mathbf{x}_n\}$ *and* $\{\boldsymbol{\varepsilon}_n\}$ *are sequences of* p*-dimensional random vectors such that* $\mathbf{y}_n = \mathbf{x}_n + \boldsymbol{\varepsilon}_n$, $\mathbf{x}_n$ *is* $\mathscr{F}_{n-1}$*-measurable and for some integer* $l > 0$, $\boldsymbol{\varepsilon}_n = \sum_{j=1}^l \boldsymbol{\varepsilon}_n(j)$, *where for* $1 \le j \le l$,

$$E\{\boldsymbol{\varepsilon}_n(j) | \mathscr{F}_{n+j-1}\} = 0, \qquad \sup_n E\{\|\boldsymbol{\varepsilon}_n(j)\|^\alpha | \mathscr{F}_{n+j-1}\} < \infty \quad a.s., \tag{A.4}$$

for some $\alpha > 2$. *Also assume that*

$$\lambda_n = \lambda_*\left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' + \sum_{i=1}^n \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i'\right) \to \infty \quad a.s., \tag{A.5}$$

$$\log \lambda^*\left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right) = o(\lambda_n) \quad a.s. \tag{A.6}$$

Then

$$\lim_{n\to\infty} \lambda_*\left(\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i'\right) \Big/ \lambda_n = 1 \quad a.s. \tag{A.7}$$

PROOF. Let $R_n = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$ and $G_n = \sum_{i=1}^n \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i'$. Then

$$\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i' = R_n + \sum_{i=1}^n \mathbf{x}_i \boldsymbol{\varepsilon}_i' + \sum_{i=1}^n \boldsymbol{\varepsilon}_i \mathbf{x}_i' + G_n. \tag{A.8}$$

We can assume that R_n is nonsingular a.s. Otherwise for $j = -p + 1, \ldots, 0$,

define $\mathbf{y}_j = \mathbf{x}_j = \mathbf{e}_j$ and $\boldsymbol{\varepsilon}_j = \mathbf{0}$, where $\{\mathbf{e}_j\}$ is an orthonormal basis. This reduces the problem to the case that R_n is nonsingular. Now, by (A.4) and Lemma 1 of Lai and Wei (1982), for each j, $\mathbf{x}_j$ is $\mathcal{F}_{n+j-1}$-measurable and

$$\left\| R_n^{-1/2} \sum_{i=1}^{n} \mathbf{x}_i \boldsymbol{\varepsilon}_i'(j) \right\|^2 = O(\log \lambda^*(R_n)) \quad \text{a.s.}$$

Consequently,

$$\left\| R_n^{-1/2} \sum_{i=1}^{n} \mathbf{x}_i \boldsymbol{\varepsilon}_i' \right\|^2 = O(\log \lambda^*(R_n)) \quad \text{a.s.} \tag{A.9}$$

Given any unit vector $\mathbf{u}$, by (A.6) and (A.9),

$$\begin{aligned}
\mathbf{u}' \sum_{i=1}^{n} \mathbf{x}_i \boldsymbol{\varepsilon}_i' \mathbf{u} &= \mathbf{u}' R_n^{1/2} \left(R_n^{-1/2} \sum_{i=1}^{n} \mathbf{x}_i \boldsymbol{\varepsilon}_i' \right) \mathbf{u} \\
&\le \left\| \mathbf{u}' R_n^{1/2} \right\| \left\| R_n^{-1/2} \sum_{i=1}^{n} \mathbf{x}_i \boldsymbol{\varepsilon}_i' \right\| \\
&\le \{\mathbf{u}'(R_n + G_n)\mathbf{u}\}^{1/2} O\left[\log^{1/2} \lambda^*(R_n)\right] \\
&\le \mathbf{u}'(R_n + G_n)\mathbf{u} O\left[\{\log \lambda^*(R_n)/\lambda_n\}^{1/2}\right] \\
&= \mathbf{u}'(R_n + G_n)\mathbf{u} o(1) \quad \text{a.s.}
\end{aligned}$$

Applying this to (A.8), we have

$$\mathbf{u}' \left(\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i' \right) \mathbf{u} = \mathbf{u}'(R_n + G_n)\mathbf{u}(1 + o(1)) \quad \text{a.s.} \tag{A.10}$$

Since the $o(1)$ term does not involve $\mathbf{u}$, (A.7) follows. □

LEMMA A.2. *If all eigenvalues of a $p \times p$ matrix A have magnitudes* 1, *then*

$$\lambda_n = \lambda_* \left(\sum_{i=1}^{n} A^i (A')^i \right) \to \infty \quad \text{as } n \to \infty. \tag{A.11}$$

PROOF. We prove (A.11) by contradiction. Since λ_n is increasing and A is nonsingular, if (A.11) does not hold, then there is $\lambda > 0$ such that $\lim_{n\to\infty} \lambda_n = \lambda$. Let $H_n = \sum_{i=1}^{n} A^i (A')^i$ and $\mathbf{e}_n$ be the eigenvector corresponding to λ_n such that $\|\mathbf{e}_n\| = 1$. Then there exists a subsequence n_j and a unit vector $\mathbf{e}$ such that $\lim_{j\to\infty} \mathbf{e}_{n_j} = \mathbf{e}$. Fix n. We have

$$\left\| H_n^{1/2} \mathbf{e} \right\| \le \left\| H_n^{1/2} (\mathbf{e} - \mathbf{e}_{n_j}) \right\| + \left\| H_n^{1/2} \mathbf{e}_{n_j} \right\|. \tag{A.12}$$

But since $n_j > n$,

$$\left\| H_n \mathbf{e}_{n_j} \right\|^2 = \mathbf{e}_{n_j}' H_n \mathbf{e}_{n_j} \le \mathbf{e}_{n_j}' H_{n_j} \mathbf{e}_{n_j} = \lambda_{n_j} \le \lambda. \tag{A.13}$$

In view of (A.12) and (A.13), for all n,

$$\left\| H_n^{1/2}\mathbf{e} \right\| \leq \lim_{j\to\infty} \left\| H_n^{1/2}(\mathbf{e} - \mathbf{e}_{n_j}) \right\| + \lambda^{1/2} = \lambda^{1/2}.$$

Consequently,

$$0 < \sum_{i=1}^{\infty} \mathbf{e}'A^i(A')^i\mathbf{e} \leq \lambda < \infty. \tag{A.14}$$

Now, by Jordan decomposition, there exists an orthonormal (possibly complex) matrix P such that

$$PAP^* = \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & D_k \end{pmatrix}, \tag{A.15}$$

where P^* is the Hermitian transpose of P and D_j are $d_i \times d_i$ matrices with $\sum_{j=1}^{k} d_j = p$ and

$$D_j = \begin{pmatrix} \lambda_j & 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & 0 \\ & & & \ddots & 1 \\ 0 & & & & \lambda_j \end{pmatrix}, \qquad |\lambda_j| = 1. \tag{A.16}$$

In view of (A.14)–(A.16), it is sufficient to show that if $\|\mathbf{u}\| = 1$ and

$$0 < \sum_{i=1}^{\infty} \mathbf{u}'D^i(D^*)^i\mathbf{u}^* < \infty \tag{A.17}$$

for some $d \times d$ matrix D which satisfies

$$D = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & 0 \\ & & & \ddots & 1 \\ 0 & & & & \lambda \end{pmatrix}, \qquad |\lambda| = 1,$$

then we have a contradiction. For this, observe that

$$D^i = \begin{pmatrix} \lambda^i & \binom{i}{1}\lambda^{i-1} & \cdots & \binom{i}{d-1}\lambda^{i-d+1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \binom{i}{1}\lambda^{i-1} \\ 0 & & & \lambda^i \end{pmatrix}, \quad \text{where } \binom{i}{l} = 0 \text{ if } i < l.$$

Therefore, if $\mathbf{u} = (u_1, \ldots, u_d)'$, then

$$\lim_{n\to\infty} \mathbf{u}'\left(\sum_{i=1}^{n} D^i (D')^i\right)\mathbf{u}^* \Big/ \left[\sum_{i=1}^{n}\binom{i}{d-1}^2\right] = |u_1|^2.$$

By (A.17), $|u_1^2| = 0$. Now if $u_1 = \cdots = u_j = 0$, by a similar argument, $u_{j+1} = 0$. Consequently, $\mathbf{u} = \mathbf{0}$, which contradicts (A.17). □

LEMMA A.3. *Let $\{\boldsymbol{\delta}_n, \mathcal{F}_n\}$ be a sequence of p-dimensional martingale differences such that for some $\alpha > 2$ and matrix Γ,*

$$E(\boldsymbol{\delta}_n \boldsymbol{\delta}_n' | \mathcal{F}_{n-1}) = \Gamma \quad \text{and} \quad \sup_n E\left(\|\boldsymbol{\delta}_n\|^\alpha \big| \mathcal{F}_{n-1}\right) < \infty \quad a.s. \tag{A.18}$$

Let $\mathbf{z}_{n+1} = A\mathbf{z}_n + \boldsymbol{\delta}_n$, where $\mathbf{z}_0$ is $\mathcal{F}_0$-measurable and A a $p \times p$ matrix. Assume that

$$\text{all eigenvalues of } A \text{ have magnitudes } 1, \tag{A.19}$$

$$\Gamma_q = \sum_{j=0}^{q} A^j \Sigma (A')^j \text{ is nonsingular for some } q > 0. \tag{A.20}$$

Then

$$\liminf_{n\to\infty} \lambda_* \left(\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i'\right)\frac{1}{n} = \infty \quad a.s. \tag{A.21}$$

PROOF. For any positive integer l,

$$\mathbf{z}_{n+1} = A^{lq+1}\mathbf{z}_{n-lq} + \sum_{j=0}^{lq} A^j \boldsymbol{\delta}_{n-j}.$$

Under assumptions (A.18)–(A.20), it is known [Lai and Wei (1985), Theorems 1 and 2] that for some integer m,

$$\lambda^*\left(\sum_{i=lq}^{n} A^{lq+1}\mathbf{z}_{i-lq}\mathbf{z}_{i-lq}' A^{lq+1}\right) = O(n^m) \quad \text{a.s.}, \tag{A.22}$$

and

$$\begin{aligned}\sum_{k=lq}^{n}\left(\sum_{j=0}^{lq} A^j \boldsymbol{\delta}_{k-j}\right)\left(\sum_{j=0}^{lq} A^j \boldsymbol{\delta}_{k-j}\right)' &\sim n\sum_{j=0}^{lq} A^j \Gamma (A')^j \\ &= n\sum_{s=1}^{l} A^s \left(\sum_{j=0}^{q} A^j \Sigma (A')^j\right)(A')^s \\ &= n\sum_{s=1}^{l} A^s \Gamma_q (A')^s \quad \text{a.s.}\end{aligned} \tag{A.23}$$

In view of (A.22) and (A.23), we can apply Lemma A.1 to obtain

$$\liminf_{n\to\infty} \lambda_*\left(\sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i'\right)\frac{1}{n} \geq \lambda_*\left(\sum_{s=1}^{l} A^s\Gamma_q(A')^s\right)$$

$$\geq \lambda_*(\Gamma_q)\lambda_*\left(\sum_{s=1}^{l} A^s(A')^s\right) \quad \text{a.s.}$$

Since l can be chosen arbitrarily, Lemma A.2 and (A.20) imply (A.21). □

PROOF OF THEOREM A.1. Let $\boldsymbol{\delta}_n = (\delta_n, 0, \ldots, 0)'$ and

$$A = \begin{pmatrix} \theta_1, \ldots, \theta_{t-1} & \theta_t \\ I_{t-1} & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} \psi_1, \ldots, \psi_{s-1} & \psi_s \\ I_{s-1} & 0 \end{pmatrix}.$$

Then all eigenvalues of A have magnitudes 1 and all eigenvalues of B have magnitudes less than 1. We also have

$$\mathbf{v}_n = A\mathbf{v}_{n-1} + \boldsymbol{\delta}_n \quad \text{and} \quad \mathbf{u}_n = B\mathbf{u}_{n-1} + \boldsymbol{\delta}_n. \tag{A.24}$$

Let $\Gamma = E(\delta_n\delta_n') = \sigma^2 M$, where $M_{11} = 1$ and other $M_{ij} = 0$. It is known [see Lai and Wei (1985), Theorems 1 and 2 and Example 3] that

$$\lambda^*(H_n) = O(n^\rho) \quad \text{for some } \rho > 0, \tag{A.25}$$

$$\lim_{n\to\infty} n^{-1}P_n = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \mathbf{u}_i\mathbf{u}_i' = \sum_{j=0}^{\infty} B^j\Gamma(B')^j \quad \text{a.s.}, \tag{A.26}$$

and

$$\sum_{j=0}^{s-1} B^j\Gamma(B')^j \text{ and } \sum_{j=0}^{t-1} A^j\Gamma(A')^j \text{ are positive definite.} \tag{A.27}$$

Therefore, to show (A.3), it is sufficient to prove

$$(nH_n)^{-1/2}\sum_{i=1}^{n} \mathbf{v}_i\mathbf{u}_i' \to 0 \quad \text{a.s. as } n \to \infty. \tag{A.28}$$

First, let us assume that for all integers l,

$$(nH_n)^{-1/2}\sum_{i=1}^{n} \mathbf{v}_i\boldsymbol{\delta}_{i+l}' \to \mathbf{0} \quad \text{a.s. as } n \to \infty. \tag{A.29}$$

Then for any integer $k > 0$,

$$\begin{aligned}(nH_n)^{-1/2}\sum_{i=k}^{n} \mathbf{v}_i\mathbf{u}_i' &= (nH_n)^{-1/2}\sum_{i=k}^{n} \mathbf{v}_i\mathbf{u}_{i-k}'(B')^k \\ &\quad + (nH_n)^{-1/2}\sum_{i=k}^{n} \mathbf{v}_j\sum_{l=0}^{k-1} \boldsymbol{\delta}_{j-l}(B')^{l-1} \\ &= I_{1n} + I_{2n} \quad \text{(say)}.\end{aligned} \tag{A.30}$$

By (A.29), $\|I_{2n}\| = o(1)$ a.s. By the Cauchy–Schwarz inequality,

$$\|I_{1n}\|^2 \le \frac{1}{n}\left(\sum_{i=k}^{n}\|H_n^{-1/2}\mathbf{v}_i\|^2\right)\left(\sum_{i=k}^{n}\|\mathbf{u}_{i-k}\|^2\right)\|B^k\|$$

$$= \operatorname{tr}\left(H_n\sum_{i=k}^{n}\mathbf{v}_i\mathbf{v}_i'\right)\left(\frac{1}{n}\sum_{i=k}^{n}\|\mathbf{u}_{i-k}\|^2\right)\|B^k\|$$

$$\le t\|B^k\|\left[n^{-1}\operatorname{tr}(P_n) - n^{-1}\|\mathbf{u}_0\|^2\right].$$

Therefore, by (A.26),

$$\text{(A.31)}\qquad \limsup_{n\to\infty}\left\|(nH_n)^{-1/2}\sum_{i=k}^{n}\mathbf{v}_i\mathbf{u}_i'\right\| \le t\|B^k\|\operatorname{tr}\left(\sum_{j=0}^{\infty}B^j\Gamma(B')^j\right).$$

Since all eigenvalues of B have magnitudes less than 1, $\lim_{k\to\infty}\|B^k\| = 0$. This and (A.31) imply (A.28).

Now, it remains to show (A.29). For $l \ge 1$, since $\mathbf{v}_i$ is $\mathcal{F}_{i+l-1}$-measurable, by Lai and Wei (1982) and (A.25),

$$\left\|H_n^{-1/2}\sum_{i=1}^{n}\mathbf{v}_i\boldsymbol{\delta}_{i+l}'\right\| = O(\log\lambda^*(H_n)) = o(\log n) \quad \text{a.s.}$$

Hence (A.27) holds. For $l \le 0$,

$$\text{(A.32)}\qquad \mathbf{v}_j = \sum_{i=0}^{-l}A^i\boldsymbol{\delta}_{j-i} + A^{1-l}\mathbf{v}_{j+l-1}$$

and

$$\text{(A.33)}\qquad \begin{aligned} H_n^{-1/2}\sum_{j=-l}^{n}\mathbf{v}_j\boldsymbol{\delta}_{j+l}' &= H_n^{-1/2}\sum_{j=-l}^{n}\left(\sum_{i=0}^{-l}A^i\boldsymbol{\delta}_{j-i}\right)\boldsymbol{\delta}_{j+l}' \\ &\quad + H_n^{-1/2}\sum_{j=-l}^{n}A^{1-l}\mathbf{v}_{j+l-1}\boldsymbol{\delta}_{j+l}' \\ &= J_{1n} + J_{2n} \quad \text{(say)}. \end{aligned}$$

Observe that by Lemma A.3 and (A.24) (with $\psi_i = 0$),

$$\text{(A.34)}\qquad \|J_{1n}\| \le \|H_n^{-1/2}\|\left(\sum_{i=0}^{-l}\|A^i\|\right)\left(\sum_{j=1}^{n}\|\boldsymbol{\delta}_j\|^2\right) = o(n^{1/2}) \quad \text{a.s.}$$

Set $R_n = A^{1-l}(\sum_{j=-l+1}^{n}\mathbf{v}_j\mathbf{v}_j')(A')^{1-l}$. Then in view of (A.9),

$$\text{(A.35)}\qquad \begin{aligned} \|J_{2n}\| &\le \|H_n^{-1/2}R_n^{1/2}\|\left\|R_n^{-1/2}\sum_{j=-l}^{n}\left(A^{1-l}\mathbf{v}_{j+l-1}\right)\boldsymbol{\delta}_{j+l}'\right\| \\ &= \|H_n^{-1/2}R_n^{1/2}\|O\left[\{\log\lambda^*(R_n)\}^{1/2}\right]. \end{aligned}$$

However, by (A.32), we can apply (A.10) to show that for any vector $\mathbf{z}$,

$$\begin{aligned}\left\|\mathbf{z}'H_n^{-1/2}R_n^{1/2}\right\|^2 &= \mathbf{z}'H_n^{-1/2}R_nH_n^{-1/2}\mathbf{z}\\ &\le \mathbf{z}'H_n^{-1/2}(H_n)H_n^{-1/2}\mathbf{z}(1+o(1))^{-1}\\ &= \|\mathbf{z}\|^2(1+o(1)).\end{aligned}$$

Consequently,

$$\left\|H_n^{-1/2}R_n^{1/2}\right\|^2 = O(1) \quad \text{a.s.}$$

In view of this, (A.25) and (A.33)–(A.35)

$$\left\|(nH_n)^{-1/2}\sum_{i=1}^{n}\mathbf{v}_i\boldsymbol{\delta}_{i+l}'\right\| = o(1) + o\left([\log n/n]^{1/2}\right) = o(1) \quad \text{a.s.}$$

This completes our proof. □

APPENDIX B

Linear processes.

LEMMA B.1. *Let $\{\delta_n\}$ be a sequence of i.i.d. random variables such that $E(\delta_1) = 0$ and $E|\delta_2|^{2r} < \infty$ for some $r \ge 2$. Assume that $Y_n = \sum_{-\infty}^{\infty} a_i\delta_{n-i}$ and $Z_n = \sum_{-\infty}^{\infty} b_i\delta_{n-i}$ with $\sum_{-\infty}^{\infty}(|a_i| + |b_i|) < \infty$. Then there exists a constant K such that for all n,*

$$E\left|\sum_{t=1}^{n}(Y_tZ_t - EY_tZ_t)\right|^r \le Kn^{r/2}.$$

A better result that only requires $\{\delta_n\}$ to be a sequence of martingale differences can be found in Findley and Wei (1990). We omit the proof of Lemma B.1.

Acknowledgments. The author would like to thank Dr. D. Findley, Dr. Hung Chen and Dr. J. Rissanen for their stimulating discussions and useful references. The author would also like to thank Mrs. Betty Vanderslice for her excellent editorial and typing work for the past 10 years. This paper was typed just before her retirement.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19** 716–723.

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

CHAN, N. H. and WEI, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.* **16** 367–401.

CHOW, Y. S. (1965). Local convergence of martingales and the law of large numbers. *Ann. Math. Statist.* **36** 552–558.

DAVIS, M. H. A. and HEMERLY, E. M. (1990). Order determination and adaptive control of ARX models using the PLS criterion. In *Proceedings of the Fourth Bad Honnef Conference on Stochastic Differential Systems. Lecture Notes in Control and Inform. Sci.* (N. Christopeit, ed.). Springer, New York.

DRAPER, N. R. and SMITH, H. (1981). *Applied Regression Analysis*, 2nd ed. Wiley, New York.

FINDLEY, D. F. and WEI, C. Z. (1990). The bias properties of AIC for possibly misspecified stationary stochastic regression model. Technical report, Dept. Mathematics, Univ. Maryland.

FRIEDLANDER, B. (1982). Lattice filters for adaptive processing. *Proceedings of the IEEE* **70** 829–867.

HALD, A. (1952). *Statistical Theory with Engineering Applications*. Wiley, New York.

HANNAN, E. J. (1987). Rational transfer function approximation. *Statist. Sci.* **2** 135–161.

HANNAN, E. J., MCDOUGALL, A. J. and POSKIT, D. S. (1989). Recursive estimation of autoregressions. *J. Roy. Statist. Soc. Ser. B* **51** 217–233.

HEMERLY, E. M. and DAVIS, M. H. A. (1989). Strong consistency of the PLS criterion for order determination of autoregressive processes. *Ann. Statist.* **17** 941–946.

HUANG, D. (1990). Selecting order for general autoregressive models by minimum description length. *J. Time Ser. Anal.* **11** 107–118.

KAVALIERIS, L. (1989). The estimation of the order of an autoregression using recursive residuals and cross-validation. *J. Time Ser. Anal.* **10** 271–281.

LAI, T. L., ROBBINS, H. and WEI, C. Z. (1979). Strong consistency of least squares estimates in multiple regression. II. *J. Multivariate Anal.* **9** 343–361.

LAI, T. L. and WEI, C. Z. (1982a). Least squares estimates in stochastic regression models with applications to identification and control systems. *Ann. Statist.* **10** 154–166.

LAI, T. L. and WEI, C. Z. (1982b). Asymptotic properties of projections with applications to stochastic regression problems. *J. Multivariate Anal.* **12** 346–370.

LAI, T. L. and WEI, C. Z. (1983). Asymptotic properties of general autoregressive models and strong consistency of least squares estimates of their parameters. *J. Multivariate Anal.* **13** 1–23.

LAI, T. L. and WEI, C. Z. (1984). Moment inequalities with applications to regression and time series models. In *Inequalities in Statistics and Probability* (Y. L. Tong, ed.) 165–172. IMS, Hayward, Calif.

LAI, T. L. and WEI, C. Z. (1985). Asymptotic properties of multivariate weighted sums with applications to stochastic regression in linear dynamic systems. In *Multivariate Analysis VI* (P. R. Krishnaiah, ed.) 375–393. North-Holland, Amsterdam.

LAI, T. L. and WEI, C. Z. (1986). On the concept of excitation in least squares identification and adaptive control. *Stochastics* **16** 227–254.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

PAULSEN, J. (1984). Order determination of multivariate autoregressive time series with unit roots. *J. Time Ser. Anal.* **5** 115–127.

PHILLIPS, P. C. B. (1986). Understanding spurious regressions in econometrics. *J. Econometrics* **33** 311–340.

PÖTSCHER, B. M. (1989). Model selection under nonstationary autoregressive models and stochastic linear regression models. *Ann. Statist.* **17** 1257–1274.

RISSANEN, J. (1986a). Stochastic complexity and modeling. *Ann. Statist.* **14** 1080–1100.

RISSANEN, J. (1986b). A predictive least squares principle. *IMA J. Math. Control Inform.* **3** 211–222.

RISSANEN, J. (1986c). Order estimation by accumulated prediction errors. In *Essays in Time Series and Allied Processes* (J. Gani and M. B. Priestly, eds.). *J. Appl. Probab.* **23A** 55–61.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 416–464.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147.

TSAY, R. S. (1984). Order selection in nonstationary autoregressive models. *Ann. Statist.* **12** 1425–1433.

WAX, M. (1988). Order selection for AR models by predictive least squares. *IEEE Trans. Acoust. Speech Signal Process.* **36** 581–588.

WEI, C. Z. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *Ann. Statist.* **15** 1667–1682.

YOKOYAMA, R. (1980). Moment bounds for stationary mixing sequences. *Z. Wahrsch. Verw. Gebiete* **52** 45–57.

ZYGMUND, A. (1959). *Trigonometric Series* **1**. Cambridge Univ. Press.

INSTITUTE OF STATISTICAL SCIENCE
ACADEMIA SINICA
TAIPEI 11529, TAIWAN
REPUBLIC OF CHINA

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND 20742