The Annals of Statistics
1989, Vol. 17, No. 3, 1257–1274

MODEL SELECTION UNDER NONSTATIONARITY: AUTOREGRESSIVE MODELS AND STOCHASTIC LINEAR REGRESSION MODELS[1]

BY B. M. PÖTSCHER

Yale University

We give sufficient conditions for strong consistency of estimators for the order of general nonstationary autoregressive models based on the minimization of an information criterion à la Akaike's (1969) AIC. The case of a time-dependent error variance is also covered by the analysis. Furthermore, the more general case of regressor selection in stochastic regression models is treated.

1. Introduction. The statistical properties of order estimation and model selection procedures based on so-called information theoretic criteria like Akaike's AIC or its variants have been intensively studied in recent years. Most of the work has been done in the field of time series analysis. In the framework of stationary autoregressive models strong consistency of the order estimators obtained through minimization of certain variants of AIC, like BIC, has been discussed in Hannan and Quinn (1979) for the univariate case, and in Quinn (1980) for the multivariate case. Parallel results for stationary autoregressive moving average models are given in Hannan (1980, 1981). The latter two papers give also weaker conditions under which weak consistency holds. For an alternative approach using sequences of tests see Pötscher (1983, 1985). Weak consistency results for the closely related model selection problem for linear regression models with asymptotically stationary regressors and normal i.i.d. errors are given in Geweke and Meese (1981) [for a review of the literature on selection of regressors, see Amemiya (1980) and Thompson (1978a, b)]. On the contrary, for nonstationary autoregressive models with i.i.d. errors weak consistency of the order estimators has been established independently by Paulsen (1984) and Tsay (1984); the former also treating the multivariate case. The nonstationarity considered in both papers arises from the fact that the characteristic polynomial is allowed to have roots not only outside but also on the unit circle. Another case of nonstationarity is considered in Paulsen and Tjøstheim (1985): Here the autoregressive scheme has to be stable, that is, all zeros of the characteristic polynomial are outside the unit circle, but the error process is allowed to have a nonconstant variance.

The present paper gives strong consistency results for model selection procedures based on variants of AIC for general nonstationary stochastic linear regression models where the errors constitute a not necessarily stationary

Received November 1986; revised October 1988.

[1] Research supported by the Max Kade Foundation.

AMS 1980 *subject classifications.* Primary 62M10, 62J05, 60G10; secondary 62F12, 93E12.

Key words and phrases. Model selection, order estimation, selection of regressors, strong consistency, autoregression, nonstationarity, nonergodic models, information criteria.

martingale difference sequence using recent results of Lai and Wei (1982a, b, 1983, 1985). These results are then applied to yield strong consistency results for order estimation in nonstationary autoregressive models. The assumptions used in the present paper are weaker than each of the assumptions employed in Paulsen (1984), Paulsen and Tjøstheim (1985) and Tsay (1984); hence they provide a common framework for models exhibiting both kinds of nonstationarity. After the first version of this paper was written, Wei brought the related papers by Wang and An (1984), An and Gu (1985) and Gu and An (1985) to my attention. The first one of these papers discusses model selection in stochastic linear regression models and autoregressive models by means of BIC, that is, criterion (2.1) with $C(T) = \log T$ in our notation. In this paper Wang and An give conditions for "overconsistency" of the selected models, that is, conditions under which the selected models contain all relevant regressors but possibly also redundant ones. Their paper neither gives conditions for the full consistency property nor explores the domain of feasible rates for the penalty term $C(T)$ such that consistency results as is done in this paper. For a further discussion of the results in Wang and An (1984) and their relation to the present paper see Sections 3 and 4. The two other papers, An and Gu (1985) and Gu and An (1985), deal essentially only with the stationary case. The paper is organized as follows: Section 2 gives consistency results in the general context of a linear regression model with stochastic regressors and martingale difference errors; in Section 3 the special case of autoregressive models is treated; Section 4 contains complementary remarks; all proofs are relegated to the Appendix.

2. Model selection in linear regression models. The dependent variable of a linear regression is modeled as a real-valued stochastic process (y_t). The family of potential regressors under consideration is a family $\mathscr{R} = ((z_{tk})\colon k \in \mathscr{K})$ of real-valued stochastic processes defined on the same probability space $(\Omega, \mathscr{F}, P)$ as y_t is. The set $\mathscr{K}$ is an arbitrary index set and the index t varies in $\mathbb{N}$, the set of positive integers. More specifically, the researcher has in mind a (nonvoid) set $\mathscr{M}$ of regression models M, where each M is a finite subset of $\mathscr{K}$, that is, under model M the regressors (z_{tk}) for $k \in M$ enter the regression equation for (y_t) (if M is void regression is on the null space). Important special cases are the case where the regressors are ordered in a natural way, for example, $\mathscr{K}$ is $\mathbb{N}$, or an initial segment of $\mathbb{N}$ (in the natural order), and M runs through all initial segments of $\mathscr{K}$, or the case where $\mathscr{M}$ is the set of all finite subsets of $\mathscr{K}$. We shall be concerned with the problem of choosing a "minimal" and "true" model M from the set $\mathscr{M}$. The notion of a true model is here to be understood in the sense of the following definition and is defined relative to a given filtration $\mathscr{F}_s$, $s \in \mathbb{N} \cup \{0\}$, of the σ-field $\mathscr{F}$. *In this section, the filtration* $\mathscr{F}_s$ *will be throughout assumed to have the property that* z_{tk} *is* $\mathscr{F}_{t-1}$*-measurable for all* $t \in \mathbb{N}$ *and* $k \in \mathscr{K}$. The prototypical examples for this situation are the case of an autoregressive model where $\mathscr{F}_{t-1}$ represents, for example, the σ-field generated by the past of the process (y_t), or the case of a general linear dynamic regression model. *For the rest of this section, we shall also assume that a true model in the sense of Definition* 2.1 *exists in* $\mathscr{M}$.

DEFINITION 2.1. A model M is called a true model for (y_t) relative to $(\mathscr{F}_t)$ if there exist real numbers β_k, $k \in M$, such that y_t can be P-a.s. decomposed as $y_t = \Sigma_{k \in M}\beta_k z_{tk} + u_t$ such that for all $t \in \mathbb{N}$

(i) $E(u_t|\mathscr{F}_{t-1}) = 0$,
(ii) u_t is measurable w.r.t. $\mathscr{F}_t$.

Notice that condition (i) in Definition 2.1 determines u_t and $\Sigma\beta_k z_{tk}$ uniquely up to null sets. Similarly, if two models M and $\overline{M}$ satisfy condition (i), then the respective residuals u_t and $\bar{u}_t$ are P-a.s. equal; especially if M is a true model so is then $\overline{M}$. Notice that in a true model the error process u_t is a martingale difference sequence (taking the conditional expectation of a random variable, it is always understood that the random variable is integrable) and that no integrability assumptions have been made for y_t or z_{tk}. Of course, it is the decomposition of y_t in a linear part plus an error and the "orthogonality" condition (i) which reflect the notion of a linear model; condition (ii) which makes the error process u_t then a martingale difference sequence is only necessary to make the asymptotics work.

The selection of a model given the first T observations of y_t and z_{tk} will be based on minimization over $\mathscr{M}$ of one of the criterion functions

$$\log \hat{\sigma}_T^2(M) + \operatorname{size}(M)C(T)/T \tag{2.1}$$

or

$$\hat{\sigma}_T^2(M) + \operatorname{size}(M)C(T)/T, \tag{2.2}$$

where $\hat{\sigma}_T^2(M)$ is the residual variance after fitting model M. Here $C(T)$ denotes a nonnegative real-valued random variable; further properties of $C(T)$ will be specified later. The quantity size(M), which is assumed to be real-valued, stands for any measure of the size or complexity of the model M as, for example, the number of parameters. If size(M) is chosen to be the number of parameters and $C(T) = 2$, then (2.1) reduces to Akaike's (1969) AIC criterion; if $C(T) = \log T$, (2.1) reduces to Schwarz's (1978) BIC criterion. Notice also that for the minimization of (2.1) it is irrelevant whether in (2.1) the term $\hat{\sigma}_T^2(M)$ is replaced by the residual sum of squares RSS(M) or not, since $\hat{\sigma}_T^2(M) = \mathrm{RSS}(M)/T$. The results of this section also apply to criteria similar to (2.1) and (2.2), where the penalty term size$(M)C(T)/T$ is replaced by $C(M,T)/T$, or where size(M) is random; see Section 4.

The following lemmas are the essential building blocks for the consistency result. The first lemma gives conditions which guarantee that the value of the criterion function (2.1) or (2.2) at an incorrect model is eventually larger than the corresponding value at a true model. We use the following notation and conventions: The quantity $\hat{\sigma}_T^2(M)$ is given as $T^{-1}y'(I - Z_M(Z_M'Z_M)^+Z_M')y$, where $y = (y_1, \dots, y_T)'$. The matrix Z_M is $T \times m$, where $m = \operatorname{card}(M)$, with its tth row equal to $(z_{tk_1}, \dots, z_{tk_m})$ and $k_1, \dots, k_m$ enumerate M. (If M is void or M contains only the zero regressor, then $Z_M = 0 \in \mathbb{R}^T$.) The projection matrix on the column space of Z_M is denoted by $P_M = Z_M(Z_M'Z_M)^+Z_M'$, where A^+ denotes the Moore–Penrose inverse of A and the prime denotes transposition. Furthermore, we use $f = (f_1, \dots, f_T)'$, $u = (u_1, \dots, u_T)'$, where f_t and u_t are defined (up to null sets) via the decomposition: $y_t = f_t + u_t$ P-a.s., f_t is $\mathscr{F}_{t-1}$-measurable and $E(u_t | \mathscr{F}_{t-1}) = 0$. Such a decomposition exists in view of Definition 2.1 and of the assumed existence of a true model (if y_t is integrable of course such a decomposition always exists). We note that f_t and u_t do not depend on a particular model and that u_t can be chosen to satisfy (ii) of Definition 2.1. As a convention we set $\log 0 = -\infty$, $\log(a/0) = \log \infty = \infty$ if $a > 0$. The symbol $\log^+ x$ is $\log x$ if $x \geq 1$ and 0 if $0 \leq x < 1$. For $x \in \mathbb{R}^T$ we set $\|x\| = (x'x)^{1/2}$. We say that a sequence E_T, $E_T \subseteq \Omega$, is eventual if for almost all $\omega \in \Omega$ the relation $\omega \in E_T$ holds for $T \geq T(\omega)$. Now consider the following condition with $M \in \mathscr{M}$:

(1) $\| f - P_M f \|^2 \to \infty$ a.s. and $[\log^+ \operatorname{tr}(Z_M' Z_M)] / \| f - P_M f \|^2$ converges to 0 a.s. as $T \to \infty$.

LEMMA 2.1. *Let* $M_1 \in \mathscr{M}$ *be a true model and* $M_2 \in \mathscr{M}$. *If* $\sup_{t \geq 1} E(|u_t|^\alpha | \mathscr{F}_{t-1}) < \infty$ *a.s. for some* $\alpha > 2$ *and if* (1) *holds for* M_2, *then we have*:

(a) $\hat{\sigma}_T^2(M_2) > 0$ *eventually, hence* $\log(\hat{\sigma}_T^2(M_2)/\hat{\sigma}_T^2(M_1))$ *is well defined eventually.*

(b) $\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1) > (\operatorname{size}(M_1) - \operatorname{size}(M_2))C(T)/T$ *holds eventually if* $C(T)/\| f - P_{M_2} f \|^2$ *converges to* 0 *a.s. as* $T \to \infty$.

(c) $\log(\hat{\sigma}_T^2(M_2)/\hat{\sigma}_T^2(M_1)) > (\operatorname{size}(M_1) - \operatorname{size}(M_2))C(T)/T$ *holds eventually if* $C(T)/(T \log(1 + \| f - P_{M_2} f \|^2 (u'u)^{-1}))$ *converges to* 0 *a.s. as* $T \to \infty$.

We note that under (1) the quantity $\| f - P_{M_2} f \|^2$ is eventually positive, hence $C(T)/\| f - P_{M_2} f \|^2$ is eventually well defined, and similarly $B(T) = \log(1 + \| f - P_{M_2} f \|^2 (u'u)^{-1})$ is eventually well defined and positive (possibly $+\infty$ if $u'u = 0$); for later use we remark that under $\sup T^{-1} u'u < \infty$ a.s. the condition $C(T)/(T \log(1 + \| f - P_{M_2} f \|^2 T^{-1})) \to 0$ a.s. implies $C(T)/TB(T) \to 0$ a.s., and the converse is true if $\liminf T^{-1} u'u > 0$ a.s. holds. (Note that $\sup E(|u_t|^\alpha | \mathscr{F}_{t-1}) < \infty$ a.s. for some $\alpha > 2$ implies $u'u = \sum_{t=1}^T E(u_t^2 | \mathscr{F}_{t-1}) + o(T)$ [see Chow (1965) and Lai and Wei (1982a)]; hence $\sup T^{-1} u'u < \infty$ a.s.) Under the classical assumptions for the asymptotic theory in linear regression models $\| f - P_{M_2} f \|^2$ and $u'u$ behave like T and then the conditions in (b) and (c) reduce to $C(T)/T \to 0$ a.s.

Notice also that eventual positivity of $\| f - P_{M_2} f \|^2$ implies that M_2 is not a true model. Condition (1) is a separation condition, that is, it tells us how well separated from the true models the incorrect models have to be in order that they can be recognized as such, cf. Remark 1. Furthermore, since all the conditions on M_2 in Lemma 2.1 are only in terms of P_{M_2} and $\operatorname{tr}(Z_{M_2}' Z_{M_2})$, they depend only on the space spanned by the model M_2, that is, the image of P_{M_2}, and not on the special way this space is represented by the regressors in M_2 (cf. Lemma A.2 in the Appendix).

The next lemma gives conditions under which in particular the value of the criterion function (2.1) or (2.2) at a true model of minimal size is eventually smaller than the corresponding value at a nonminimal true model. We introduce the condition (2) for a model $M \in \mathscr{M}$:

$$C(T) \to \infty \text{ a.s. and } [\log^+ \operatorname{tr}(Z_M' Z_M)]/C(T) \to 0 \text{ a.s. as } T \to \infty. \tag{2}$$

Lemma 2.2. *Let $M_1 \in \mathscr{M}$ and $M_2 \in \mathscr{M}$ be true models with* $\operatorname{size}(M_1) < \operatorname{size}(M_2)$. *If* $\sup_{t \geq 1} E(|u_t|^\alpha | \mathscr{F}_{t-1}) < \infty$ *a.s. for some $\alpha > 2$ and if M_1 and M_2 satisfy* (2), *then we have*:

(a) $\hat{\sigma}_T^2(M_i) > 0$ *eventually* [*hence* $\log(\hat{\sigma}_T^2(M_2)/\hat{\sigma}_T^2(M_1))$ *is eventually well defined*] *and* $\log(\hat{\sigma}_T^2(M_2)/\hat{\sigma}_T^2(M_1)) > (\operatorname{size}(M_1) - \operatorname{size}(M_2))C(T)/T$ *eventually holds, if* $\liminf_{T \to \infty} T^{-1} u'u > 0$ *a.s. and* $[\log^+ \operatorname{tr}(Z_{M_i}' Z_{M_i})]/T \to 0$ *a.s. as $T \to \infty$ for $i = 1, 2$.*

(b) $\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1) > (\operatorname{size}(M_1) - \operatorname{size}(M_2))C(T)/T$ *holds eventually.*

Clearly, the a.s. boundedness of $C(T)/T$ is a sufficient condition for $[\log^+ \operatorname{tr}(Z_{M_i}' Z_{M_i})]/T \to 0$ a.s. under (2). Notice that under $\sup E(|u_t|^\alpha | \mathscr{F}_{t-1}) < \infty$ a.s. for some $\alpha > 2$ the condition $\liminf T^{-1} u'u > 0$ a.s. is equivalent to $\liminf_{T \to \infty} T^{-1} \Sigma_{t=1}^T E(u_t^2 | \mathscr{F}_{t-1}) > 0$ a.s., which is in turn implied by $\liminf E(u_t^2 | \mathscr{F}_{t-1}) > 0$ a.s., a condition used in Lai and Wei (1982b, 1983). Furthermore, condition (2) and the conditions in (a) depend on the models M_1, M_2 only through the space spanned by them [note, however, that models spanning the same space may have assigned different values of the complexity measure $\operatorname{size}(M)$].

A simple consequence of Lemmas 2.1 and 2.2 is Theorem 2.3. Denote by $\hat{M}(T, 1)$ and $\hat{M}(T, 2)$, respectively, an *arbitrary* model which minimizes the criterion function (2.1) or (2.2), respectively, over $\mathscr{M}$. Under a minimal true model we understand a true model which has minimal size among all true models.

Theorem 2.3. *Let $\mathscr{M}$ be finite and assume* $\sup_{t \geq 1} E(|u_t|^\alpha | \mathscr{F}_{t-1}) < \infty$ *a.s. for some $\alpha > 2$.*

(a) *Assume that for each $M \in \mathscr{M}$, which is not a true model, condition* (1) *holds. Then*:

(a.1) $\hat{M}(T, 1)$ *is a true model eventually if*

$$C(T) / \left(T \log\left(1 + \|f - P_M f\|^2 (u'u)^{-1}\right)\right)$$

goes to 0 *a.s. as $T \to \infty$ for all $M \in \mathscr{M}$ which are not true models.*

(a.2) $\hat{M}(T, 2)$ *is a true model eventually if* $C(T)/\|f - P_M f\|^2$ *goes to* 0 *a.s. as $T \to \infty$ for all $M \in \mathscr{M}$ which are not true models.*

(b) *Assume that for each true model $M \in \mathscr{M}$ condition* (2) *holds. Then*:

(b.1) $\hat{M}(T, 1)$ *is a minimal true model or a false model eventually if* $\liminf_{T \to \infty} T^{-1} u'u > 0$ *a.s. and if* $[\log^+ \operatorname{tr}(Z_M' Z_M)]/T \to 0$ *a.s. as $T \to \infty$ for all true $M \in \mathscr{M}$.*

(b.2) $\hat{M}(T, 2)$ *is a minimal true model or a false model eventually.*

Combining parts (a) and (b) of Theorem 2.3, one obtains rather general conditions under which the selected models $\hat{M}(T,1)$ and $\hat{M}(T,2)$ are eventually minimal and correct. Note that we have not assumed in Theorem 2.3 that only one minimal true model exists.

Remark 1. (i) The conditions involving $\| f - P_M f \|^2$ and $\operatorname{tr}(Z_M' Z_M)$ in the results above can be expressed in terms of eigenvalues. For example the condition $\lim \| f - P_M f \|^2 = \infty$ a.s. for a model M which is not a true model is implied by $\lim \lambda_{\min}((Z_M\colon f)'(Z_M\colon f)) = \infty$ a.s. in view of (1.6) in Lai and Wei (1982b). This simply means that f and Z_M are asymptotically not multicollinear. The other condition in (1) essentially balances the growth rates of the largest eigenvalue of $Z_M' Z_M$ and of the smallest eigenvalue of $(Z_M\colon f)'(Z_M\colon f)$. The conditions on $C(T)$ in Lemma 2.1 give, loosely speaking, an upper bound for the growth rate of $C(T)$ in terms of the growth rate of $\lambda_{\min}((Z_M\colon f)'(Z_M\colon f))$. Similarly, the conditions on $C(T)$ in (2) typically specify a minimal rate of divergence of $C(T)$ in terms of $\lambda_{\max}(Z_M' Z_M)$, M now a true model.

(ii) In the special case where all models $M \in \mathscr{M}$ are submodels of an overall true model (not necessarily a member of $\mathscr{M}$), that is, Z_M is a selection of columns from a matrix X corresponding to the overall model, then the condition $\lim \lambda_{\min}(X'X) = \infty$ a.s. is sufficient for $\lim \| f - P_M f \|^2 = \infty$ a.s., where M is not a true model: Clearly, $\| f - P_M f \|^2 \geq \| f - P_{M^*} f \|^2$, where Z_{M^*} is a matrix obtained from X by deleting a column which is not a column of Z_M and which has a nonzero coefficient β_k in the overall model. This choice is possible since M is not a true model. But then $\| f - P_{M^*} f \|^2 \geq \beta_k^2 K^{-1} \lambda_{\min}(X'X)$, where K is the number of columns of X. Similarly, given $\lim \lambda_{\min}(X'X) = \infty$ a.s., it follows for incorrect models M that $\lambda_{\min}^{-1}(X'X) \log \lambda_{\max}(X'X) \to 0$ a.s. implies the second part of (1) and that the conditions for $C(T)$ in Lemma 2.1 are now satisfied if $C(T)/\lambda_{\min}(X'X) \to 0$ a.s. and $C(T)/(T \log(1 + \lambda_{\min}(X'X)(u'u)^{-1})) \to 0$ a.s., respectively; furthermore, condition (2) is then implied by $\log \lambda_{\max}(X'X)/C(T) \to 0$ a.s. (Of course an overall true model can always be constructed; however the sufficient conditions in terms of $X'X$ may then be overly restrictive.)

Remark 2. Lemma 2.1 [and hence Theorem 2.3(a)] also holds without the conditional moment condition if condition (1) is replaced by the following condition: $\| f - P_{M_2} f \|^2 > 0$ eventually and $\| f - P_{M_2} f \|^{-2} u'u \to 0$ a.s. as $T \to \infty$. This condition may be useful in a situation where the second part of (1) is violated. We note that the proof of this version of Lemma 2.1 does not make use of (ii) of Definition 2.1, that is, of $\mathscr{F}_t$-measurability of u_t, and of $\mathscr{F}_{t-1}$-measurability of z_{tk} (if z_{tk} is not $\mathscr{F}_{t-1}$-measurable, P-a.s. $\mathscr{F}_{t-1}$-measurability of $\sum_{k \in M} \beta_k z_{tk}$ has then to be added as a condition to Definition 2.1). Furthermore, Lemmas 2.1 and 2.2 and Theorem 2.3 hold for $\alpha = 2$ if some of the other assumptions are slightly sharpened. For details see Pötscher (1986).

Remark 3. Inspection of the proof of Lemma 2.2 shows that the following more general result holds. Let $M_1 \in \mathscr{M}$ and $M_2 \in \mathscr{M}$ be true models with $\operatorname{size}(M_1) < \operatorname{size}(M_2)$. If (i) $u'u > 0$ eventually, (ii) $(u'u)^{-1} u' P_{M_i} u \to 0$ a.s.,

(iii) $TC(T)^{-1}(u'u)^{-1}u'P_{M_i}u \to 0$ a.s. and (iv) $C(T) > 0$ eventually, then the conclusions of Lemma 2.2(a) hold. Clearly, (i), (ii) and a.s. boundedness of $T/C(T)$ imply (i)–(iv). Another set of sufficient conditions for (i)–(iv), which will be of importance in Section 3, is the following: $C(T)^{-1}u'P_{M_i}u \to 0$ a.s., $T^{-1}u'P_{M_i}u \to 0$ a.s., $\liminf T^{-1}u'u > 0$ a.s. and $C(T) > 0$ eventually. Similarly, if $C(T)^{-1}u'P_{M_i}u \to 0$ a.s. and $C(T) > 0$ eventually, then the conclusion of Lemma 2.2(b) holds. (In fact, the proof of Lemma 2.2 proceeds by verifying the latter two sets of conditions.) Of course, the condition $C(T)^{-1}u'u \to 0$ a.s. suffices for $C(T)^{-1}u'P_{M_i}u \to 0$ a.s. Although in many cases where Lemma 2.2(b) applies the condition $C(T)^{-1}u'u \to 0$ a.s. will give a weaker result, it may be of value if $\log^+ \operatorname{tr}(Z'_{M_i}Z_{M_i})$ increases faster than $u'u$. Finally, a similar remark on the measurability conditions for u_t and z_{tk} as in Remark 2 applies to the proofs of the versions of Lemma 2.2 discussed here.

REMARK 4. We note that the assumption $\liminf T^{-1}u'u > 0$ a.s., which means that the noise does not die out eventually, has only been used in the "underestimation" part related to criterion function (2.1). If the magnitude of the errors does not affect the regressors, for example, if no feedback is present, then clearly an (asymptotically) vanishing noise should make model selection easier and therefore a condition like $\liminf T^{-1}u'u > 0$ a.s. is unnatural in this context. (If feedback is present as in the case of an autoregression, then a vanishing noise entails a degenerate design matrix, thus possibly compensating the positive effect; cf. Remark 4, Section 3.) As the above results show we can indeed avoid explicit use of this condition for the criterion function (2.2). For the criterion function (2.1) the "overestimation" part also works without this condition since the systematic bias in $\hat{\sigma}_T^2(M_2)$, M_2 not a true model but satisfying (1), dominates the residual error variance $\hat{\sigma}_T^2(M_1)$ of a true model anyway. However, in the "underestimation" part we have to differentiate between different true models M_1 and M_2 on the basis of $\log \hat{\sigma}_T^2(M_1)$, $\log \hat{\sigma}_T^2(M_2)$ and their respective sizes if we use criterion (2.1). Now if the noise is for example not existent, that is, $u'u = 0$ a.s., then the first term in (2.1) equals $-\infty$ regardless of the size of M_1, M_2 and hence we cannot discriminate between a minimal and a nonminimal true model in this case. This is of course an extreme case and the condition $\liminf T^{-1}u'u > 0$ can be relaxed somewhat, cf. Remark 3.

The following example shows how Theorem 2.3 works in the framework of asymptotically stationary processes. It also shows how asymptotic stationarity can be used to relax the condition on the penalty term $C(T)$ resulting from Theorem 2.3.

EXAMPLE. Let z_{tk}, $1 \le k \le K$, $K \in \mathbb{N}$, be jointly asymptotically stationary processes in the sense that $T^{-1}\Sigma_{t=1}^{T} z_{tk}z_{tj}$ converges a.s. to some real-valued (possibly stochastic) quantity q_{kj}. We assume that the processes are not multicollinear in the sense that the matrix $Q = (q_{kj})$, $1 \le j$, $k \le K$, is a.s. positive definite. The process y_t is generated as $y_t = \Sigma_{k=1}^{K}\beta_k z_{tk} + u_t$, where u_t is a martingale difference sequence w.r.t. a filtration $\mathscr{F}_t$ such that z_{tk} is $\mathscr{F}_{t-1}$-measurable (typically the filtration will be generated by current and past errors, current and past regressors as well as the regressors leading by one). We assume that $\sup E(|u_t|^{\alpha}|\mathscr{F}_{t-1}) < \infty$ a.s. for some $\alpha > 2$ and that $\liminf T^{-1}u'u > 0$ a.s. If we choose $\mathscr{M}$ as the set of all models M_l, $0 \le l \le K$, where M_l contains all regressors with indices running from 1 to l, and M_0 is the void model, we face the situation of model selection where we have an a priori ordering of the regressors expressed through their enumeration. If we choose $\mathscr{M}$ as the set of all possible models (including the void model) with regressors from $(z_{tk}, 1 \le k \le K)$, then we have the subset selection problem. In the first case we want to end up with the model M_{l_0}, where l_0 is the maximum of $\{k\colon \beta_k \ne 0\}$ and $l_0 = 0$ if this set is empty; in the second case the desired model contains only the regressors with indices from the set $\{k\colon \beta_k \ne 0\}$. We shall now verify the conditions of Theorem 2.3 and show that in both cases we shall eventually pick the desired model if either one of the model selection criteria (2.1) or (2.2) is used with size(M) = number of regressors in M [size$(M) = 0$ if M is void] and $C(T)$ is such that $\log T/C(T)$ and $C(T)/T$ converge to 0 a.s. First of all a model M is true iff it contains all regressors with indices from $\{k\colon \beta_k \ne 0\}$. One-half of this statement is trivial; the other one follows from the fact that $\liminf T^{-1}\|f - P_M f\|^2 \ge \liminf T^{-1}\|f - P_{M^*} f\|^2 \ge \beta_{k_1}^2 K^{-1}\lambda_{\min}(Q) > 0$ a.s., where M^* contains all variables z_{tk} except for $k = k_1$ where $\beta_{k_1} \ne 0$ and z_{tk_1} is not contained in M (if $y_t = u_t$, then there are only true models and the above claim is trivial). Hence we have also shown that $\|f - P_M f\|^2$ grows at least linearly for M not a true model. Next for an arbitrary model M we have $\lim T^{-1}\operatorname{tr}(Z_M' Z_M) = \Sigma q_{kk}$ a.s., where the summation is over all k such that $k \in M$, hence $\log^{+}\operatorname{tr}(Z_M' Z_M) = O(\log T)$ a.s. But then all conditions in Theorem 2.3 are satisfied (note that $\sup T^{-1}u'u < \infty$ a.s.) which gives the desired conclusion. The condition $\log T/C(T) \to 0$ a.s. can be weakened to $\log\log T/C(T) \to 0$ a.s. if we additionally assume that $z_{Tk}^2 = o(T^{\gamma})$ a.s. for some $0 < \gamma < 1$ and all $1 \le k \le K$. To verify this claim, we have, in the light of Remark 3, only to show that $C(T)^{-1}u'P_M u \to 0$ a.s. for any true model M. This is trivial for the void model, hence assume $M \ne \varnothing$. Since $Z_M' Z_M/T$ converges now a.s. to a nonsingular matrix by assumption and since $\Sigma_{t=1}^{T} z_{tk}u_t = O(T^{1/2}(\log\log T)^{1/2})$ a.s. for $1 \le k \le K$ by Lemma 2 in Wei (1985), the result follows. Obviously in this example the analogous results hold for any other set of models $\mathscr{M}$ too.

3. Autoregressive models. In this section we apply the general results obtained in Section 2 to the order estimation problem in general autoregressive models. We shall first discuss nonexplosive processes and purely explosive ones and then general autoregressive processes. The result for nonexplosive processes will then be sharpened in the case of stable autoregressive processes.

Let us now fix the notation. The process y_t is for $t \ge 1$ assumed to be generated by

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_{p_0} y_{t-p_0} + u_t, \tag{3.1}$$

where $\beta = (\beta_1, \dots, \beta_{p_0})'$ is the parameter vector satisfying $\beta_{p_0} \neq 0$. Let $\mathscr{F}_0$ denote the σ-field generated by the starting values $y_0, \dots, y_{1-p_0}$ (if $p_0 = 0$, i.e., $y_t = u_t$, then $\mathscr{F}_0$ is set equal to the trivial σ-field) and let $\mathscr{F}_t$ be the σ-field generated by the starting values and $u_1, \dots, u_t$. The error process is assumed to be a martingale difference with respect to the filtration $\{\mathscr{F}_t\}$, that is, $E(u_t|\mathscr{F}_{t-1}) = 0$. These assumptions will be maintained throughout this section, except where otherwise noted.

The order p_0 of the autoregressive process will in general not be known and hence has to be estimated. One way to do this is to minimize as usual one of the criterion functions (2.1) or (2.2) over all autoregressive models M_p of order p, $0 \leq p \leq P$, where it is assumed that $p_0 \leq P$ and $P \geq 1$ is a prespecified constant. In more detail, autoregressive models of order p are fitted to the data $y_{P+1}, \dots, y_T$ by least squares and the residual variance $\hat{\sigma}_T^2(p)$ is calculated which is then used to calculate (2.1) and (2.2). We set $y = (y_{P+1}, \dots, y_T)'$, $u = (u_{P+1}, \dots, u_T)'$, $f = (f_{P+1}, \dots, f_T)'$ and Z_p is the $(T - P) \times p$ matrix whose ith row is given by $(y_{P-1+i}, \dots, y_{P-p+i})$. If $p = 0$ we put $Z_p = 0 \in \mathbb{R}^{T-P}$. This conforms with the notation of the previous section if we take into account that the sample period used for the calculation of $\hat{\sigma}_T^2(p)$ is now $P + 1 \leq t \leq T$, that is, $\hat{\sigma}_T^2(p) = (T - P)^{-1}y'(I - Z_p(Z_p'Z_p)^+Z_p')y$. Let $\hat{p}_T(1)$ and $\hat{p}_T(2)$, respectively, denote a minimizer of $\log \hat{\sigma}_T^2(p) + pC(T)/T$ and of $\hat{\sigma}_T^2(p) + pC(T)/T$ over $0 \leq p \leq P$, respectively. Then we have the following consistency result for nonexplosive processes.

THEOREM 3.1. *Assume that the characteristic polynomial of* (3.1), *that is,* $1 - \beta_1 z - \cdots - \beta_{p_0} z^{p_0}$, *has all its zeros outside or on the unit circle in the complex plane. Let* $\sup_{t \geq 1} E(|u_t|^\alpha|\mathscr{F}_{t-1}) < \infty$ *a.s. for some* $\alpha > 2$ *and* $\liminf_{T\to\infty} T^{-1}\Sigma_{t=1}^T E(u_t^2|\mathscr{F}_{t-1}) > 0$ *a.s. hold. Then* $\hat{p}_T(1) \geq p_0$ *and* $\hat{p}_T(2) \geq p_0$ *hold eventually if* $C(T)/T \to 0$ *a.s. as* $T \to \infty$. *Furthermore,* $\hat{p}_T(1) \leq p_0$, $\hat{p}_T(2) \leq p_0$ *eventually if* $C(T)/\log T \to \infty$ *a.s. as* $T \to \infty$.

Of course combining both parts of the theorem shows that a growth rate of $C(T)$ between $\log T$ and T gives consistent estimators. For purely explosive processes Theorem 3.2 provides the analogous result. Note that the assumptions of Theorem 3.2 and $\beta_{p_0} \neq 0$ imply that (3.1) is purely explosive in the sense of Lai and Wei (1983).

THEOREM 3.2. *Assume that the characteristic polynomial of* (3.1) *has all its zeros inside the unit circle in the complex plane and* $p_0 \geq 1$ *holds. Assume* $\sup_{t \geq 1} E(|u_t|^\alpha|\mathscr{F}_{t-1}) < \infty$ *a.s. for some* $\alpha > 2$ *and* $\liminf_{t\to\infty} E(u_t^2|\mathscr{F}_{t-1}) > 0$ *a.s. hold. Then* $\hat{p}_T(1) \geq p_0$ *holds eventually if* $C(T)/T^2 \to 0$ *a.s. as* $T \to \infty$. *Similarly,* $\hat{p}_T(2) \geq p_0$ *holds eventually if* $C(T)/e^{aT} \to 0$ *a.s. as* $T \to \infty$ *for some* $0 < a = a(\omega) < -2\log m$, *where m is the maximum of the moduli of the zeros of the characteristic polynomial. If* $\liminf_{T\to\infty} C(T)/T > 0$ *a.s., then* $\hat{p}_T(1) \leq p_0$ *and* $\hat{p}_T(2) \leq p_0$ *eventually.*

The slightly stronger assumption on the conditional variance of the error process in Theorem 3.2 compared to Theorem 3.1 is needed to ensure the validity of Theorem 2 in Lai and Wei (1983) which gives an exponential rate for the eigenvalues of $Z_{p_0}'Z_{p_0}$. The general case, that is, the case where there are no restrictions on the location of the zeros of the characteristic polynomials, is treated in the next theorem.

THEOREM 3.3. *Assume* $\sup_{t\geq 1} E(|u_t|^{\alpha}|\mathscr{F}_{t-1}) < \infty$ *a.s. for some* $\alpha > 2$ *and* $\liminf_{t\to\infty} E(u_t^2|\mathscr{F}_{t-1}) > 0$ *a.s. hold. Then* $\hat{p}_T(1) \geq p_0$ *and* $\hat{p}_T(2) \geq p_0$ *hold eventually if* $C(T)/T \to 0$ *a.s. as* $T \to \infty$. *Furthermore,* $\hat{p}_T(1) \leq p_0$ *and* $\hat{p}_T(2) \leq p_0$ *hold eventually if* $\liminf_{T\to\infty} C(T)/T > 0$ *a.s.*

REMARK 1. It is interesting to note that (2.1) and (2.2) lead to different feasible rates for $C(T)$ in case of explosive processes. Notice also that Theorem 3.3 does not give a common feasible rate for $C(T)$ such that both parts of that theorem are satisfied and consistency is ensured. Of course such a rate for $C(T)$ may nevertheless exist since Theorem 3.3 gives sufficient conditions only. A proof of such a result, however, would essentially have to produce a convergence rate for the least-squares estimator in a general autoregressive model which seems to be very difficult. The consistency proof for general autoregressive models in Lai and Wei (1983) may be a starting point for such a result.

REMARK 2. The proof of the overestimation part (i.e., of the first half) of Theorem 3.3 is based on the method of proof used in Wang and An (1984). The proofs of the underestimation parts of Theorems 3.2 and 3.3 are based on generalizations of Lemma 2.2 discussed in Remark 3, Section 2, and on Lemma A.1. Lemma A.1 is implicit in the proof of Theorem 1 of Lai and Wei (1983) and appears also in Wang and An (1984). In the context of general autoregressive models Wang and An (1984) prove only that $\hat{p}_T(1) \geq p_0$ eventually a.s. for $C(T) = \log T$ (they actually treat the corresponding problem of subset selection of autoregressive parameters, see also Remark 3 below).

REMARK 3. The results of this section can be extended to model selection in autoregressions where models are chosen from a more general set of models $\mathscr{M}$, to some extent. Theorems 3.1 and 3.4 carry over completely to this case; however the full strength, for example, of the overestimation part of Theorem 3.2, does not necessarily go through. This is so because estimating $\| f - P_M f \|^2$ for wrong models M from below by $\lambda_{\min}(Z_{\overline{M}}'Z_{\overline{M}})$, where $\overline{M}$ is the smallest true model containing M, may give only a linear growth rate since the difference equation corresponding to model $\overline{M}$ is not necessarily purely explosive in the sense of Lai and Wei (1983) although the true process is. Another potential difficulty is that the proof of Theorem 3.3 relies on the consistency of least-squares estimators in general autoregressive models as established in Lai and Wei (1983). If a subset autoregression is estimated by least squares, then this consistency result does not immediately apply. [Referring to the result of Lai and Wei (1983), Wang and An (1984) make use of consistency of the least-squares estimator in general subset autoregressions in the proof of their result mentioned in Remark 2; An has informed me that in the meantime he has found a proof of this consistency result.]

REMARK 4. (i) The conditions $\liminf T^{-1}\Sigma E(u_t^2|\mathscr{F}_{t-1}) > 0$ a.s. and $\liminf E(u_t^2|\mathscr{F}_{t-1}) > 0$ a.s. are, in the context of autoregressive models, used to ensure that the design matrix Z_p does not degenerate and that the lower bounds for $\lambda_{\min}(Z_p'Z_p)$ given in Lai and Wei (1983, 1985) hold. Hence these conditions are—in contrast to Remark 4 in Section 2—now essential for the *overestimation* parts of the theorems of this section. As discussed in Remark 4 in Section 2 a condition of this type is also used to prove the underestimation parts for the estimator $\hat{p}_T(1)$. In contrast to that the underestimation result for $\hat{p}_T(2)$ in Theorem 3.1 and in Theorems 3.2 and 3.3 [in the latter two theorems under the slightly stronger condition $C(T)/T \to \infty$ a.s.] hold without any of these two conditions as can be seen from the proofs. Furthermore, $\hat{p}_T(2) \le p_0$ eventually holds under the single condition $u'u/C(T) \to 0$ a.s. [and $C(T) > 0$ eventually] without any further assumptions on u_t, cf. Remark 3 in Section 2. However, without such further conditions also models with $p < p_0$ might be true models.

(ii) Under the assumptions of the theorems of this section p_0 is of course uniquely determined. Clearly, this is already true under weaker conditions on u_t.

Finally, we show that in the stable case the condition $C(T)/\log T \to \infty$ can be weakened to $C(T)/\log\log T \to \infty$. A similar result was proved in Hannan and Quinn (1979) in the stationary case [and where $\hat{\sigma}_T^2(p)$ was obtained from the Yule–Walker equations]. We have not been able to prove such a result also for nonexplosive models although it might be possible.

THEOREM 3.4. *Assume that the characteristic polynomial of* (3.1) *has all its zeros outside the unit circle in the complex plane*, $\sup_{t\ge 1} E(|u_t|^\alpha|\mathscr{F}_{t-1}) < \infty$ *a.s. for some* $\alpha > 2$, *and that* $\liminf_{T\to\infty} T^{-1}\Sigma_{t=1}^T E(u_t^2|\mathscr{F}_{t-1}) > 0$ *a.s. holds. Then* $\hat{p}_T(1) \ge p_0$, $\hat{p}_T(2) \ge p_0$ *eventually hold if* $C(T)/T \to 0$ *a.s. as* $T \to \infty$, *and* $\hat{p}_T(1) \le p_0$, $\hat{p}_T(2) \le p_0$ *eventually hold if* $C(T)/\log\log T \to \infty$ *a.s. as* $T \to \infty$.

Under stronger assumptions on u_t, the $\log\log T$ bound for $C(T)$ in Theorem 3.4 can be slightly weakened to $\liminf C(T)/\log\log T > c$ for a suitable constant c. This bound is then sharp, in general; compare the discussion in Hannan and Quinn (1979), page 193. We note that Akaike's AIC, that is, (2.1) with $C(T) = 2$ and $\text{size}(M_p) = p$, does not satisfy all the conditions in the above theorems. Actually, AIC does not even give weakly consistent order estimators for nonexplosive processes, see Tsay (1984). The criterion BIC, that is, (2.1) with $C(T) = \log T$ and $\text{size}(M_p) = p$, does not satisfy all the conditions in Theorem 3.1. A close inspection of the proofs of Lemma 2.2 and Theorem 3.1, however, reveals that Theorem 3.1 is still valid if $C(T)/\log T \to \infty$ a.s. is replaced by $\liminf C(T)/\log T > c$ a.s. The constant c, however, depends on (y_t). Thus if BIC would be modified to BIC′ by setting $C(T) = c'\log T$, $c' > c$, then we could be sure that BIC′ gives consistent estimates. This is of course of no great practical interest since c is unknown. Furthermore, BIC gives consistent estimators under the conditions of Theorem 3.4 of course. Finally, we note that the theorems in this section obviously remain true if the upper bound P increases with the sample size (possibly depending on ω) slowly enough. We do not know how fast P is allowed to increase. For some results in this direction for stationary autoregressions, see, for example, An, Chen and Hannan (1982).

4. Complementary remarks. The "overconsistency" result for stochastic linear regression models selected by BIC as given in Wang and An (1984) can be derived under weaker assumptions as is seen from an inspection of their proof: Conditions (2.6) and (2.8) can be dropped without any loss since—in the notation of that paper—$S(J_k)^{-1}$ can be estimated from below by $(s's)^{-1}$ and hence Lemma 2.1 in that paper is not necessary for the proof of their Theorem 2.1 [notice that their assumption (2.1) implies $(s's)^{-1} > MT^{-1}$, $M = M(\omega) > 0$].

The results of Section 2 easily carry over to more general criteria of the form (2.1) or (2.2), where the penalty term size$(M)C(T)/T$ is replaced by $C(M, T)/T$. Lemma 2.1 carries over where now the conditions for $C(T)$ in (b) and (c) of this lemma have to be satisfied by $\Delta C(M_1, M_2, T) = C(M_1, T) - C(M_2, T)$. For Lemma 2.2 to carry over the penalty term has to be such that at least for true models M and M' always either $\Delta C(M, M', T) > 0$, $= 0$ or < 0 eventually holds. This gives then the required ordering of the models according to their "size." Lemma 2.2 is then true if all conditions for $C(T)$ are satisfied by $-\Delta C(M_1, M_2, T)$ and $\Delta C(M_1, M_2, T) < 0$ eventually holds. These results are of some importance for an analysis of model selection criteria such as Mallows' (1973) C_p, or in a situation where $C(M, T) = \text{size}(M)C(T)$ but size(M) is random.

APPENDIX

Proof of Lemma 2.1. We start from the basic a.s. identity

$$\begin{aligned} T\big(\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1)\big) & \\ = \|f - P_{M_2} f\|^2 + 2f'(I - P_{M_2})u - u'P_{M_2}u + u'P_{M_1}u. & \end{aligned} \tag{A.1}$$

Under our assumptions the second term on the r.h.s. of (A.1) is

$$O\Big(\|f - P_{M_2} f\|\big[\max\big(1, \log^+(\|f - P_{M_2} f\|), \log^+\big(\operatorname{tr}(Z'_{M_2}Z_{M_2})\big)\big)\big]^{1/2}\Big) \quad \text{a.s.}$$

and the term $u'P_{M_2}u$ on the r.h.s. of (A.1) is

$$O\big(\big\{\max\big[1, \log^+\big(\operatorname{tr}(Z'_{M_2}Z_{M_2})\big)\big]\big\}\big) \quad \text{a.s.}$$

This follows from Lai and Wei (1982b), Theorems 4 and 3 [in Theorem 3, (2.2), a printing error occurs: The term in braces on the r.h.s. of (2.2) should read $\max(1, \log^+(\Sigma\Sigma z_{ij}^2))$]. Since $\|f - P_{M_2} f\| \to \infty$ a.s. under (1) and since the last term on the r.h.s. of (A.1) is nonnegative, we hence have for every $0 < \varepsilon < 1$

eventually

$$\text{(A.2)} \qquad \hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1) > \varepsilon\| f - P_{M_2} f \|^2/T.$$

This proves parts (a) and (b).

Since

$$\hat{\sigma}_T^2(M_1) = T^{-1}\big(u'u - u'P_{M_1}u\big) \le T^{-1}u'u$$

we obtain from (A.2) under either set of assumptions that eventually

$$\text{(A.3)} \qquad \big[\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1)\big]\hat{\sigma}_T^{-2}(M_1) \ge \varepsilon\| f - P_{M_2} f \|^2 (u'u)^{-1}.$$

Notice that (A.3) trivially holds if $\hat{\sigma}_T^2(M_1) = 0$ or $u'u = 0$ because $\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1) > 0$ eventually as shown above. Now the r.h.s. of (A.3) is eventually positive. But then

$$\begin{aligned} \log\big(\hat{\sigma}_T^2(M_2)/\hat{\sigma}_T^2(M_1)\big) &= \log\big(1 + \big[\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1)\big]\hat{\sigma}_T^{-2}(M_1)\big) \\ &\ge \log\Big(1 + \varepsilon\| f - P_{M_2} f \|^2 (u'u)^{-1}\Big) \end{aligned}$$

eventually and hence (c) holds, since $\log(1 + x)/\log(1 + \varepsilon x)$ is bounded on the interval $(0, \infty)$. □

Proof of Lemma 2.2. Clearly, for $i = 1, 2$ we have $T\hat{\sigma}_T^2(M_i) = u'u - u'P_{M_i}u$ a.s. From Lai and Wei (1982b), Theorem 3, we get

$$u'P_{M_i}u = O\Big(\Big\{\max\Big[1, \log^+\Big(\operatorname{tr}\big(Z'_{M_i}Z_{M_i}\big)\Big)\Big]\Big\}\Big)$$

a.s. Now since $C(T) \to \infty$ and $[\log^+ \operatorname{tr}(Z'_{M_i}Z_{M_i})]/C(T) \to 0$ a.s. under (2) we get $\lim C(T)^{-1}u'P_{M_i}u = 0$ a.s. Since $\operatorname{size}(M_1) < \operatorname{size}(M_2)$ this proves (b) because

$$TC(T)^{-1}\big(\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1)\big) = C(T)^{-1}\big(u'P_{M_1}u - u'P_{M_2}u\big)$$

which goes to 0 as just shown. Now under the assumptions of (a) we clearly have $\liminf \hat{\sigma}_T^2(M_i) = \liminf T^{-1}u'u > 0$ a.s. since $\lim T^{-1}u'P_{M_i}u = 0$ a.s. follows from $[\log^+ \operatorname{tr}(Z'_{M_i}Z_{M_i})]/T \to 0$ and Theorem 3 in Lai and Wei (1982b). This clearly implies $\hat{\sigma}_T^2(M_i) > 0$ eventually; $\lim(\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1))\hat{\sigma}_T^{-2}(M_1) = 0$ a.s. also follows. Finally,

$$\log\big(\hat{\sigma}_T^2(M_2)/\hat{\sigma}_T^2(M_1)\big) = \log\big(1 + \big(\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1)\big)\hat{\sigma}_T^{-2}(M_1)\big)$$

and $(\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1))\hat{\sigma}_T^{-2}(M_1)$ goes to zero as just shown. Hence

$$\begin{aligned} &TC(T)^{-1}\log\big(\hat{\sigma}_T^2(M_2)/\hat{\sigma}_T^2(M_1)\big) \\ &\quad = (1 + \xi_T)^{-1}TC(T)^{-1}\big(\hat{\sigma}_T^2(M_2) - \hat{\sigma}_T^2(M_1)\big)\hat{\sigma}_T^{-2}(M_1), \end{aligned}$$

where ξ_T is a mean value going to 0 a.s. But then (a) follows since the r.h.s. of the last equation goes to 0 a.s. by what has already been established. □

Proof of Theorem 2.3. Since $\mathscr{M}$ is finite $\hat{M}(T, 1)$ and $\hat{M}(T, 2)$ exist. We give the proof for (a.1) and (b.1) only; the proof for (a.2) and (b.2) is completely analogous. Assume (a.1) is not true. Since $\mathscr{M}$ is finite we would have $\hat{M}(T, 1) = M_2$

infinitely often on a set of positive probability, where M_2 is a fixed model which is not true. Choose a fixed true model $M_1 \in \mathscr{M}$. Applying Lemma 2.1 to M_1 and M_2, we see that the value of the criterion function (2.1) at M_2 is eventually larger than the value at M_1 which leads to a contradiction. To prove (b.1), assume that it would not be true. Then similar as above $\hat{M}(T, 1) = M_2$ infinitely often with positive probability, where M_2 is now a fixed nonminimal true model. By finiteness of $\mathscr{M}$ a minimal true model exists, say M_1. Now from Lemma 2.2 we conclude that the value of (2.1) at M_2 is eventually larger than the value at M_1, which leads to a contradiction. □

In the following proofs we use results from Section 2. Recall that the actual sample size in Section 3 is $T - P$; hence, in order to apply the results of Section 2 properly the penalty term $C(T)$ has to be translated into a penalty term $C'(T - P)$, where $C'(T - P) = C(T)(T - P)/T$. Since all conditions for $C(T)$ used below hold for $C(T)$ iff they hold for $C'(T - P)$ we shall ignore this difference for the sake of brevity.

Proof of Theorem 3.1. First we show that an autoregressive model M_p with $p < p_0$ is not true by showing that $\liminf_{T \to \infty}(T - P)^{-1}\| f - P_{M_p} f \|^2 > 0$ a.s. Of course, it suffices to prove this for $p = p_0 - 1$. Now a.s.

$$(T - P)^{-1}\| f - P_{M_{p_0-1}} f \|^2 = (T - P)^{-1}\| Z_{p_0}\beta - P_{M_{p_0-1}} Z_{p_0}\beta \|^2$$

$$= (T - P)^{-1}\| v\beta_{p_0} - P_{M_{p_0-1}} v\beta_{p_0} \|^2,$$

where v denotes the last column of Z_{p_0}. The last expression equals

$$\beta_{p_0}^2 (T - P)^{-1}\| v - P_{M_{p_0-1}} v \|^2 \geq p_0^{-1}\beta_{p_0}^2 (T - P)^{-1}\lambda_{\min}\left(Z_{p_0}' Z_{p_0}\right),$$

the inequality following from (1.6) in Lai and Wei (1982b). Since $\beta_{p_0} \neq 0$ and $\liminf(T - P)^{-1}\lambda_{\min}(Z_{p_0}' Z_{p_0}) > 0$ a.s. by Theorem 3 and Example 3 in Lai and Wei (1985), we arrive at the desired conclusion. Furthermore, for $p < p_0$ we have

$$\log^+ \operatorname{tr}\left(Z_p' Z_p\right) \leq \log^+ \operatorname{tr}\left(Z_{p_0}' Z_{p_0}\right) \leq \log^+ p_0 + \log^+ \lambda_{\max}\left(Z_{p_0}' Z_{p_0}\right).$$

It follows now from Corollary 1 in Lai and Wei (1985) that $\lambda_{\max}(Z_{p_0}' Z_{p_0}) = O(T)$ a.s. if all zeros of the characteristic polynomial are outside the unit circle and $\lambda_{\max}(Z_{p_0}' Z_{p_0}) = O(T^{2\rho} \log\log T)$ a.s. otherwise where ρ is the sum of the multiplicities of all of the zeros on the unit circle. In any case $\log^+ \lambda_{\max}(Z_{p_0}' Z_{p_0}) = O(\log T)$ a.s. Since $\| f - P_{M_p} f \|^2$ increases at least as T as just shown above, we have $\log^+ \operatorname{tr}(Z_{p_0}' Z_{p_0})/\| f - P_{M_p} f \|^2 \to 0$ a.s. for $p < p_0$. This shows that assumption (1) is satisfied and, applying Theorem 2.3, we obtain the first half of Theorem 3.1 taking into account the remarks after Lemma 2.1. On the other hand it is obvious that M_p is a true model if $p \geq p_0$. Furthermore,

$$\log^+ \operatorname{tr}\left(Z_p' Z_p\right) \leq \log^+ p + \log^+ \lambda_{\max}\left(Z_p' Z_p\right) = O(\log T) \quad \text{a.s.},$$

again by Corollary 1 in Lai and Wei (1985). Hence $\log^+ \operatorname{tr}(Z_p' Z_p)/T \to 0$ a.s. and $\log^+ \operatorname{tr}(Z_p' Z_p)/C(T) \to 0$ a.s. since $C(T)/\log T \to \infty$ a.s. by assumption. This establishes condition (2) and the conditions in Theorem 2.3. The theorem then follows from Theorem 2.3. □

Proof of Theorem 3.2. For $p < p_0$ we have similarly as before

$$\| f - P_{M_p} f \|^2 \geq \| f - P_{M_{p_0-1}} f \|^2 \geq p_0^{-1}\beta_{p_0}^2 \lambda_{\min}\left(Z_{p_0}' Z_{p_0}\right) \geq p_0^{-1}\beta_{p_0}^2 e^{aT}$$

eventually for $0 < a < -2\log m$ using Theorem 2 in Lai and Wei (1983) (to make this theorem applicable, the origin of time has to be shifted). Since $\log^+ \operatorname{tr}(Z_p' Z_p) \leq \log^+ \operatorname{tr}(Z_{p_0}' Z_{p_0}) = O(T)$ a.s. in view of Corollary 1 in Lai and Wei (1985), this shows that condition (1) is satisfied. Furthermore, $C(T)/e^{aT} \to 0$ a.s. clearly implies $C(T)/\| f - P_{M_p} f \|^2 \to 0$ a.s. and $C(T)/T^2 \to 0$ a.s. implies

$$C(T)/T \log\left(1 + \| f - P_{M_p} f \|^2 (u'u)^{-1}\right) \to 0 \quad \text{a.s.}$$

since certainly $\sup T^{-1}u'u < \infty$ a.s. The first half of the theorem then follows from Theorem 2.3. Now for $p \geq p_0$ the model M_p is certainly true and $C(T)^{-1}u'P_{M_p}u \to 0$ a.s. by Lemma A.1 and since $\liminf C(T)/T > 0$ a.s.; this gives

$$\begin{aligned} &TC(T)^{-1}\left(\hat{\sigma}_T^2(p_2) - \hat{\sigma}_T^2(p_1)\right) \\ &\quad = T(T - P)^{-1}C(T)^{-1}\left(u'P_{M_{p_1}}u - u'P_{M_{p_2}}u\right) \to 0 \text{ a.s.} \end{aligned}$$

for $p_2 > p_1 \geq p_0$, hence $\hat{\sigma}_T^2(p_2) + p_2 C(T)/T > \hat{\sigma}_T^2(p_1) + p_1 C(T)/T$ eventually. But then $\hat{p}_T(2) \leq p_0$ eventually follows. To prove $\hat{p}_T(1) \leq p_0$ eventually observe that $\liminf T^{-1}u'u > 0$ a.s. holds and Lemma A.1 implies $\liminf \hat{\sigma}_T^2(p_1) > 0$ a.s. and $(\hat{\sigma}_T^2(p_2) - \hat{\sigma}_T^2(p_1))\hat{\sigma}_T^{-2}(p_1) \to 0$ a.s. Proceeding as in the proof of Lemma 2.2, we get $\log \hat{\sigma}_T^2(p_2) + p_2 C(T)/T > \log \hat{\sigma}_T^2(p_1) + p_1 C(T)/T$ eventually from which the result follows. [We note that, if $C(T)$ satisfies $C(T)/T \to \infty$ a.s., then $\hat{p}_T(2) \leq p_0$ eventually can be proved without Lemma A.1 by directly verifying (2) using Corollary 1 in Lai and Wei (1985).] □

Proof of Theorem 3.3. For $p < p_0$ we have $\liminf (T - P)^{-1}\| f - P_{M_p} f \|^2 > 0$ a.s. by a similar argument as in the proof of Theorem 3.1, hence M_p is true iff $p \geq p_0$. The proof of the second part of the theorem is identical to the proof of the corresponding part of Theorem 3.2. The first part is proved as follows: For $p < p_0$ we have

$$\hat{\sigma}_T^2(p) - \hat{\sigma}_T^2(p_0) \geq \hat{\sigma}_T^2(p_0 - 1) - \hat{\sigma}_T^2(p_0) = (T - P)^{-1}\hat{\beta}_{p_0}^2 \| v - P_{M_{p_0-1}} v \|^2,$$

where v is the last column of Z_{p_0} and $\hat{\beta}_{p_0}$ is the least-squares estimator for β_{p_0} based on model M_{p_0}. This follows from a standard formula used in stepwise regression and (3.3) in Lai and Wei (1982b). Hence

$$\hat{\sigma}_T^2(p) - \hat{\sigma}_T^2(p_0) \geq (T - P)^{-1} p_0^{-1}\hat{\beta}_{p_0}^2 \lambda_{\min}\left(Z_{p_0}' Z_{p_0}\right)$$

and the r.h.s. is eventually larger than $(p_0 - p)C(T)/T$ since $\hat{\beta}_{p_0}^2 \to \beta_{p_0}^2 > 0$ a.s. by Theorem 1 in Lai and Wei (1983), $\liminf T^{-1}\lambda_{\min}(Z_{p_0}' Z_{p_0}) > 0$ a.s. by Theorem 3 in Lai and Wei (1985) and since $C(T)/T \to 0$ a.s. by assumption. This shows also that

$$\liminf \hat{\sigma}_T^2(p) > \liminf\left[\hat{\sigma}_T^2(p) - \hat{\sigma}_T^2(p_0)\right] > 0 \quad \text{a.s.;}$$

furthermore, since $\hat{\sigma}_T^2(p_0) \leq (T - P)^{-1}u'u$ and $\sup T^{-1}u'u < \infty$ a.s. we obtain that $[\hat{\sigma}_T^2(p) - \hat{\sigma}_T^2(p_0)]\hat{\sigma}_T^{-2}(p_0) \geq c(\omega) > 0$ from which we conclude that $\log(\hat{\sigma}_T^2(p)/\hat{\sigma}_T^2(p_0)) > (p_0 - p)C(T)/T$ eventually holds. The result $\hat{p}_T(1) \geq p_0$, $\hat{p}_T(2) \geq p_0$ then follows along the lines of the proof of Theorem 2.3. □

Proof of Theorem 3.4. We have to prove the second half only, since the first half follows from Theorem 3.1. In light of Remark 3 in Section 2 it suffices to show that $C(T)^{-1}u'P_{M_p}u \to 0$ a.s. for $p \geq p_0$. This is trivial if $p = p_0 = 0$, hence assume $p > 0$. Since $u'P_{M_p}u = u'Z_p(Z_p'Z_p)^{+}Z_p'u$ and since Theorem 3 in Lai and Wei (1985) shows that $\liminf T^{-1}\lambda_{\min}(Z_p'Z_p) > 0$ a.s., we are finished if we can establish that $\|u'Z_p\| = O((T \log\log T)^{1/2})$ a.s.: Since $\alpha > 2$ we obtain from Theorem 1 in Lai and Wei (1985) that $y_T^2 = o(T^{\gamma})$ a.s. for some $0 < \gamma < 1$, and hence $y_{T-i}^2 = o((\Sigma_{t=P+1}^{T} y_{t-i}^2)^{\gamma})$, $1 \leq i \leq p$, because of $\liminf T^{-1}\lambda_{\min}(Z_p'Z_p) > 0$ a.s. Since $\operatorname{tr}(Z_p'Z_p) = O(T)$ a.s. from Corollary 1 in Lai and Wei (1985), we get $\|u'Z_p\| = O((T \log\log T)^{1/2})$ a.s. from Lemma 2 in Wei (1985). □

Lemma A.1. *In the notation of Section* 3 *let* $\sup_{t\geq 1} E(|u_t|^{\alpha}|\mathscr{F}_{t-1}) < \infty$ *a.s. for some* $\alpha > 2$ *and* $\liminf_{t\to\infty} E(u_t^2|\mathscr{F}_{t-1}) > 0$ *a.s. hold and let* y_t *be generated by* (3.1). *Then* $\lim_{T\to\infty} T^{-1}u'P_{M_p}u = 0$ *a.s. for* $p \geq p_0$ *holds, where* $u = (u_{P+1}, \ldots, u_T)'$.

Proof. The case $p = p_0 = 0$ is trivial. For $p \geq 1$ it is shown in the proof of Theorem 1 in Lai and Wei (1983) that there is a matrix A_T which is eventually a.s. nonsingular such that $A_T Z_p' Z_p A_T'$ converges to an a.s. nonsingular matrix. Furthermore, it is shown that $A_T Z_p' u = o(T^{1/2})$ [note that the cases $r = 0$ or $s = 0$ in the notation of Lai and Wei (1983) are covered by their arguments]. The lemma follows now from $u'P_{M_p}u = u'Z_p A_T'(A_T Z_p' Z_p A_T')^{-1}A_T Z_p' u$. □

Lemma A.2. *Let* A, B *be real matrices with countable infinitely many rows and* k_1, *respectively* k_2, *many columns. For* $T \geq 1$ *let* A_T, B_T *denote the submatrices consisting of the first* T *rows. If the column spaces of* A_T *and* B_T *coincide for all* $T \geq 1$, *then* $c_1 \operatorname{tr}(A_T'A_T) \leq \operatorname{tr}(B_T'B_T) \leq c_2 \operatorname{tr}(A_T'A_T)$ *holds for all* $T \geq 1$, *where* c_1 *and* c_2 *are positive real numbers.*

Proof. From the assumptions we immediately see that A and B span the same space. Hence we can find matrices M and N such that $AM = B$ and $BN = A$ hold. But then $A_T M = B_T$ and $B_T N = A_T$. We arrive at

$$\operatorname{tr}(B_T'B_T) = \operatorname{tr}(M'A_T'A_T M) \leq \lambda_{\max}(MM')\operatorname{tr}(A_T'A_T) = c_2 \operatorname{tr}(A_T'A_T)$$

and

$$\operatorname{tr}(A_T'A_T) \le \lambda_{\max}(NN')\operatorname{tr}(B_T'B_T) = c_1^{-1}\operatorname{tr}(B_T'B_T).$$

Clearly, $c_1 > 0$ and $c_2 > 0$, if not $A = 0$ and $B = 0$. If $A = 0$ and $B = 0$, then the result trivially holds. □

Acknowledgments. I want to thank H. Z. An, C. Z. Wei, a referee and the Associate Editor for helpful comments and C. Z. Wei for drawing my attention to the papers by H. Z. An, L. Gu and S. R. Wang. Thanks are also due to J. Wan for the translation of the paper by S. R. Wang and H. Z. An, and to B. Amato for typing the manuscript.

REFERENCES

AKAIKE, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21** 243–247.

AMEMIYA, T. (1980). Selection of regressors. *Internat. Econom. Rev.* **21** 331–354.

AN, H. Z., CHEN, Z. G. and HANNAN, E. J. (1982). Autocorrelation, autoregression and autoregressive approximation. *Ann. Statist.* **10** 926–936.

AN, H. Z. and GU, L. (1985). On the selection of regression variables. *Acta Math. Appl. Sinica* **2** 27–36.

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

GEWEKE, J. and MEESE, R. (1981). Estimating regression models of finite but unknown order. *Internat. Econom. Rev.* **22** 55–70.

GU, L. and AN, H. Z. (1985). Statistical analysis of subset AR models. *Acta Math. Appl. Sinica* **8** 433–445. (In Chinese.)

HANNAN, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8** 1071–1081.

HANNAN, E. J. (1981). Estimating the dimension of a linear system. *J. Multivariate Anal.* **11** 458–473.

HANNAN, E. J. and QUINN, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41** 190–195.

LAI, T. L. and WEI, C. Z. (1982a). Least squares estimates in stochastic regression models with applications to identification and control of dynamic 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. (1985). Asymptotic properties of multivariate weighted sums with application to stochastic regression in linear dynamic systems. In *Multivariate Analysis VI* (P. R. Krishnaiah, ed.) 375–393. North-Holland, Amsterdam.

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.

PAULSEN, J. and TJØSTHEIM, D. (1985). Least squares estimates and order determination procedures for autoregressive processes with a time dependent variance. *J. Time Ser. Anal.* **6** 117–133.

PÖTSCHER, B. M. (1983). Order estimation in ARMA-models by Lagrangian multiplier tests. *Ann. Statist.* **11** 872–885.

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

PÖTSCHER, B. M. (1986). Model selection under nonstationarity: Autoregressive models and stochastic linear regression models. Mimeo, Dept. Statistics, Yale Univ.

QUINN, B. G. (1980). Order determination for multivariate autoregression. *J. Roy. Statist. Soc. Ser. B* **42** 182–185.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

THOMPSON, M. L. (1978a). Selection of variables in multiple regression. I. *Internat. Statist. Rev.* **46** 1–20.

THOMPSON, M. L. (1978b). Selection of variables in multiple regression. II. *Internat. Statist. Rev.* **46** 129–146.

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

WANG, S. R. and AN, H. Z. (1984). Consistency of selection of variables in stochastic regression. *J. Engrg. Math.* **1** 13–22. (In Chinese.)

WEI, C. Z. (1985). Asymptotic properties of least-squares estimates in stochastic regression models. *Ann. Statist.* **13** 1498–1508.

INSTITUT FÜR ÖKONOMETRIE, OR UND SYSTEMTHEORIE
TECHNISCHE UNIVERSITÄT WIEN
ARGENTINIERSTRASSE 8/119
A-1040 WIEN
AUSTRIA