A CONSISTENT VARIABLE SELECTION CRITERION FOR LINEAR MODELS WITH HIGH-DIMENSIONAL COVARIATES

Xiaodong Zheng and Wei-Yin Loh

Utah State University and University of Wisconsin

Abstract: We consider the variable selection problem in regression models when the number of covariates is allowed to increase with the sample size. An approach of Zheng and Loh (1995) for the fixed design situation is extended to the case of random covariates. This yields a unified consistent selection criterion for both random and fixed covariates. By using $t$-statistics to order the covariates, the method requires much less computation than an all-subsets search. An application to autoregressive model selection with increasing order is given. The theory is supported by some simulation results.

Key words and phrases: Autoregressive processes, random covariates, $t$-statistic.

1. Introduction

Consider the linear regression model

$$y_i = x_{i,1} \beta_1 + \cdots + x_{i,M_n} \beta_{M_n} + \epsilon_i, \ i = 1, \ldots, n$$

which relates a response variable $y$ to a set of covariates $\{x_1, \ldots, x_{M_n}\}$. A frequently encountered problem is the selection of a subset of the covariates to keep in the final model. While it is important not to omit any true covariates (i.e., those whose population regression coefficients are nonzero), it is well known that inclusion of non-true or nuisance covariates generally reduces model prediction accuracy. Another consequence of selecting only the true covariates is that the complexity of the final regression model is reduced. This is especially desirable when the total number $M_n$ of available covariates is large. A good general discussion of the effect of variable selection on parameter estimation and prediction is given in Miller (1990). Let

$$\Gamma = \{i \mid \beta_i \neq 0, 1 \leq i \leq M_n\}$$

be the index set of the true covariates (also called the true model). Then the variable selection problem is equivalent to estimating $\Gamma$, which in this article is assumed to be independent of $n$. 
When the covariates are assumed to be nonrandom, many selection criteria have been proposed and studied. These include the FPE criterion (Thompson (1978), Shibata (1984), Zhang (1992)), cross-validation (Burman (1989), Zhang (1993), Shao (1993)) and bootstrap methods (Efron (1983), Zheng and Loh (1994)). Although the FPE criterion is inconsistent, it has the advantage (compared to the resampling methods) of being easy to use and fast to compute. A class of consistent criteria that retain the simple form of the FPE is developed in Zheng and Loh (1995).

In this paper, we consider the variable selection problem when the covariates are random. Random covariates arise in many regression applications where the values of the covariates can only be observed and are not controllable. The importance of variable selection in this case is clearly recognized by Breiman and Spector (1992), who argue that models with random covariates typically have substantially higher prediction errors than the fixed design counterparts and hence more is gained by variable selection. Other distinctions between the two different models can be found in Thompson (1978).

The traditionally recommended criterion for random design models is the $S_p$ method (Hocking (1976), Thompson (1978), Linhart and Zucchini (1986)). Let $\Theta$ be the set of the indices of variables in a particular model. Then the $S_p$ method selects the model $\hat{\Gamma}_{S_p}$ which minimizes

$$S_p(\Theta) = \frac{(n - |\Theta| - 2)}{(n - |\Theta| - 1)} \frac{1}{RSS(\Theta)}$$

over all submodels $\Theta \subseteq \Omega = \{1, \ldots, M_n\}$. Here $|\Theta|$ is the cardinality of $\Theta$ and $RSS(\Theta)$ is the residual sum of squares for model $\Theta$ fitted by the least squares method. The justification for the $S_p$ criterion is that, under joint normality of covariates and the regression error, $n^{-1}(n - 2)S_p(\Theta)$ is an unbiased estimator of the expected square prediction error for model $\Theta$. In spite of this, the estimator $\hat{\Gamma}_{S_p}$ is generally not consistent for the true model $\Gamma$ in the sense that

$$\lim_{n \to \infty} P(\hat{\Gamma}_{S_p} = \Gamma) \neq 1 \quad (3)$$

(see Breiman and Freedman (1983) for a different conclusion under the setting when the true model contains infinitely many covariates). Although certain statisticians are aware of this deficiency of the $S_p$ criterion, a rigorous proof has not been given in the literature. We give a sketch of it in the Appendix. The proof also shows that the $S_p$ criterion is able to eliminate underfitting but not overfitting models. A similar conclusion can be drawn for the FPE criterion, which estimates the true model by minimizing

$$FPE(\Theta) = RSS(\Theta) + \lambda|\Theta|\hat{\sigma}^2, \quad \Theta \subseteq \Omega, \quad (4)$$
where $\lambda$ is a positive constant and

$$\hat{\sigma}^2 = (n - M_n)^{-1}\text{RSS}(\Omega)$$

(5)

is the usual estimate of the error variance $\sigma^2$ based on the full model.

To search for a viable solution, we study the class of criteria proposed in Zheng and Loh (1995) and show that they remain consistent. This provides a unified consistent variable selection approach to both the fixed and the random design situations. Our method is developed in Section 2, where we first consider the simple situation when the covariates are pre-ordered such that the true covariates are indexed before the nuisance covariates. It is then generalized to the unordered case by use of $t$-statistics to order the covariates. In both instances, we allow $M_n$ to grow with the sample size $n$. This flexibility is important since in many applications the number of covariates is usually not small relative to $n$; (see e.g., Huber (1981) and Bickel and Freedman (1983)).

We apply our method to estimation of the true dimension of an autoregressive (AR) process in Section 3. Our results permit the selection process to include increasing-dimensional AR models. The latter is useful in practice because the true dimension is often unknown. Some simulation results to support the asymptotic theory are reported in the last section, followed by an appendix containing all the technical details.

2. A Class of Consistent Criteria

The inconsistency of the $S_p$ and FPE criteria is not unusual. A similar phenomenon also exists for the FPE selection criterion in fixed design regression (Zhang 1992). As argued by Zheng and Loh (1995), the major source of the FPE’s deficiency lies in the insufficient amount of penalty it places on $\lambda|\Theta|\hat{\sigma}^2$ for model complexity. They show that replacement of $\lambda|\Theta|\hat{\sigma}^2$ in (4) by a penalty term of the form $h_n(|\Theta|)\hat{\sigma}^2$ leads to a consistent selection criterion.

We now extend this approach to regression model (1) with high dimensional random covariates. Although modification of the $S_p$ criterion could be an alternative, the current approach has the advantage of providing a unified variable selection criterion that is consistent for both random and fixed design cases. Another feature of our method is that, by using the regression $t$-statistics to order the covariates, one only needs to search $M_n$ instead of all $2^{M_n}$ subsets of covariates for the true model. The computational savings are substantial when there are large numbers of covariates.

To fix some additional notation, rewrite model (1) as $y_i = x'_i \beta + \epsilon_i$, $i = 1, \ldots, n$, where $x'_i = (x_{i1}, \ldots, x_{i,M_n})$ and $\beta' = (\beta_1, \ldots, \beta_{M_n})$, the dependence on $n$ being understood. For any submodel $\Theta$ let $\beta_{\Theta}$ be the sub-vector of $\beta$ with components $\beta_k$, $k \in \Theta$. Similarly, we let $x_{i,\Theta}$ and $X_{\Theta} = (x_{1,\Theta}, \ldots, x_{n,\Theta})'$...
denote the corresponding $i$th design sub-vector and design matrix respectively. Note that the projection matrix $P_\Theta = X_\Theta(X_\Theta'X_\Theta)^{-1}X_\Theta'$ is invariant for any generalized inverse $(X_\Theta'X_\Theta)^{-1}$. Throughout, we drop the subscript $\Theta$ whenever $\Theta = \Omega$, the full model.

The following minimum assumptions will be in effect in this section.

(A1) $x_1, \ldots, x_n$ are i.i.d. with finite second moment and are independent of $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$, which are i.i.d. with mean 0 and variance $\sigma^2$.

(A2) $E(x_1x_1')$ is nonsingular. (This implies that $E(x_1, \Theta x_1')$ is nonsingular for every $\Theta$.)

Condition (A2) is necessary to ensure that $\beta$ uniquely minimizes the expected square error function $f(b) = E(y_i - x_i'b)^2$.

2.1. The pre-ordered case

Consider first the special case when the covariates are pre-ordered such that the true covariates are indexed before the nuisance covariates. That is, the true $\beta$ in (1) takes the form $\beta' = (\beta_1, \ldots, \beta_{k_0}, 0, \ldots, 0)$ for some $k_0$ independent of $n$ and $\beta_i \neq 0$, $i \leq k_0$. Then searching for the true model $\Gamma$ defined by (2) is equivalent to estimating the unknown $k_0$. For ease of notation, we shall denote $\Theta_k = \{1, \ldots, k\}$ and $\Theta_0 = \emptyset$. Note that $|\Theta_k| = k$. We also write $\text{RSS}_k$ for $\text{RSS}(\Theta_k)$ and $P_k$ for $P_{\Theta_k}$.

Define

$$\hat{k}_0 = \arg \min_{0 \leq k \leq M_n} \{\text{RSS}_k + h_n(k)\hat{\sigma}^2\},$$

where $\hat{\sigma}^2$ is defined in (5) and $h_n(k)$ is a nonnegative function.

We impose the following conditions on $h_n$ and $M_n$.

(B1) $M_n/n \to 0$ as $n \to \infty$.

(B2) $h_n(k)$ is nondecreasing in $k$ with $h_n(0) = 0$ and $\liminf_n h_n(k+1)/h_n(k) > 1$ for any $k \geq 1$.

(B3) For each $k \geq 1$, $h_n(k)/M_n \to \infty$ as $n \to \infty$.

(B4) For each $k \geq 1$, $n^{-1}h_n(k) \to 0$ as $n \to \infty$.

Theorem 1. Under conditions (A1), (A2) and (B1)–(B4), $\hat{k}_0 \to_p k_0$ as $n \to \infty$.

Remark.

1. The two most important conditions here are (B3) and (B4), which act in opposite directions. The appropriate growth rate of $h_n(k)$ is between $M_n$ and $n$.

Thus the number of covariates plays a critical role, in the sense that a heavier penalty function $h_n$ is required when there are many nuisance covariates. Note also that the existence of $h_n$ satisfying (B3) and (B4) is guaranteed by (B1).
2. As a direct application of Theorem 1, we have that the BIC criterion (with 
$h_n(k) = k \log n$) is consistent if $M_n = o(\log n)$. Furthermore, if $M_n = \o(\log \log n)$ then setting $h_n(k) = k \log \log n$ in (6) leads to consistency of the 
$\phi$ criterion of Hannan and Quinn (1979). On the other hand, if $M_n$ is large 
compared to $n$, the performance of these two criteria can be inadequate and 
a heavier penalty is called for; see the simulation results in Section 4.

3. Because $M_n$ may increase with $n$, the assumptions of Theorem 1 are not suffi-
cient for the asymptotic existence of $(X'X)^{-1}$ unless some additional moment 
conditions are imposed (see Section 2.2). However, it is always true that 
$\text{tr}(P) = \text{rank}(X) \leq M_n$ a.s. On the other hand, since $X'k_0Xk_0$ has fixed di-
mension $k_0$, it is of full rank for large $n$ a.s. Thus Theorem 1 allows for the 
presence of collinear nuisance covariates.

2.2. The general case

For the general case when the regression covariates are not pre-ordered, a 
natural approach is to first order them consistently and then apply criterion (6).
An appealing property of this approach is that it involves much less computation 
than an all-subsets search.

Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_{M_n})'$ be the least squares estimators based on the full 
model and let 
$$T_i = \hat{\sigma}^{-1}\hat{\beta}_i \{i\text{th diagonal element of } (X'X)^{-1}\}^{-1/2}, \ i = 1, \ldots, M_n,$$
be the corresponding $t$-statistics. We assume the following conditions, which 
imply that with probability tending to one, $(X'X)^{-1}$ exists and therefore $\hat{\beta}$ is 
unique. Similar conditions are used by Mammen (1993) to study the asymp-
totic behavior of bootstrap linear contrasts in increasing-dimension linear mod-
els.

(C1) The minimum eigenvalue $\kappa_n$ of $E(x_1x_1')$ is bounded away from zero, i.e.,
$$\kappa_n \geq \kappa > 0 \text{ for some } \kappa > 0.$$
(C2) For some $\eta > 0$, $n^{-1}M^{1+\eta} \to 0$ and 
$$\sup_n \sup_{\|d\|=1} E|d'[E(x_1x_1')]^{-1/2}x_1|^{[2/\eta]} < \infty, \tag{7}$$
where $[2/\eta]$ is the smallest integer greater than or equal to $2/\eta$.

Note that the smaller $M_n$ is, the weaker the moment condition (7) becomes. 
Further, if $x_1$ has a zero mean normal distribution as assumed in Thompson 
(1978), then $d'[E(x_1x_1')]^{-1/2}x_1$ is a standard normal random variable and $\eta$ can 
be chosen to be arbitrarily small.

The proposed variable selection procedure goes as follows:
Step 1. Sort the covariates in order of decreasing absolute values of the $t$-statistics:

$$|T_{i_1}| \geq |T_{i_2}| \geq \cdots \geq |T_{i_{M_n}}|.$$  

Step 2. Apply criterion (6) to the ordered covariates $i_1, i_2, \ldots, i_{M_n}$. That is, estimate the true model $\Gamma$ by

$$\hat{\Gamma} = \{i_1, \ldots, i_{\hat{k}^*}\},$$

where $\hat{k}^* = \arg\min_{0 \leq k \leq M_n} \{\text{RSS}^*(k) + h_n(k)\hat{\sigma}^2\}$ and $\text{RSS}^*(k)$ is the residual sum of squares for model $\{i_1, \ldots, i_k\}$.

The ordering procedure using $t$-statistics was proposed, under the name "$t_{K,i}$-directed search", in Daniel and Wood (1980), Chapter 6 as a possible tool for reducing computation in stepwise regression. However, no theoretical justification for consistency was given there. For a discussion of model selection after ordering of covariates by $t$-tests, see An and Gu (1985). The following theorem is a generalization of a theorem of Zheng and Loh (1995) from the fixed design case with a fixed number of covariates to the random design situation with the number of covariates allowed to depend on $n$.

**Theorem 2.** Suppose conditions (C1)–(C2) and the assumptions of Theorem 1 hold. Then

$$\lim_{n \to \infty} P(\min_{i \in \Gamma} |T_i| > \max_{i \notin \Gamma} |T_i|) = 1$$

and criterion (8) is consistent for $\Gamma$, i.e., $\lim_n P(\hat{\Gamma} = \Gamma) = 1$.

3. Autoregressive Model Selection

The proposed criterion in Section 2.1 can be applied to model selection in time series. We shall discuss this application in the framework of autoregressive processes.

Let $\{y_i, 1 - M_n \leq i \leq n\}$ be an autoregressive process of order $M_n$ (AR($M_n$)) satisfying

$$y_i = \beta_1 y_{i-1} + \cdots + \beta_{M_n} y_{i-M_n} + \epsilon_i, \ i = 1, \ldots, n,$$

where $\epsilon_i$ are i.i.d. with mean zero and variance $\sigma^2$. Assume that the true model is AR($k_0$), that is, $\beta_{k_0} \neq 0, \beta_j = 0, j > k_0$ and $k_0$ is independent of $n$.

The literature on the subject of estimating $k_0$ is extensive. See, for example, Akaike (1974), Hannan and Quinn (1979), Hannan (1980), Rissanen (1986) and Wei (1992). Choi (1992) gives a comprehensive survey. All existing results require a good guess of an upper bound $M \geq k_0$ and a search for the true model over AR($p$), $0 \leq p \leq M$. They also treat $M$ as fixed. This formulation is not very
practical, however. Since $k_0$ is unknown, it is conceivable that the upper bound should be larger than a small fraction of $n$. To guarantee that $M \geq k_0$, it is therefore necessary and more natural to consider the problem in such a way that $M = M_n$ may increase with $n$. We give a consistent theory for this case here. The result depends on the study of increasing-dimensional AR models.

Rewrite model (10) as $y_i = \mathbf{x}_i' \beta + \epsilon_i$, where $\mathbf{x}_i' = (y_{i-1}, \ldots, y_{i-M_n})$. Then all the previous notation applies. For instance, $\hat{k}_0$ is defined in (6), where $\text{RSS}_k$ is the residual sum of squares of the AR($k$) model fitted by the least-squares method. To avoid some technicalities, we also assume that the process \{y_i\} is stationary and normally distributed with zero mean. Stationarity implies that $x_1, \ldots, x_n$ have the same distribution, that the characteristic equation in $z$

$$h(z) = 1 - \sum_{j=1}^{k_0} \beta_j z^j = 0$$

has all $k_0$ roots outside the unit circle in the complex plane, and that the autocorrelation $\rho_{k-j} = \rho_{j-k}$ between $y_k$ and $y_j$ satisfies

$$|\rho_{k-j}| \leq Ca^{k-j}$$

for some constants $C > 0$ and $0 < a < 1$ depending only on $k_0$ and $\beta_{k_0}$ (Box and Jenkins (1976), Section 3.2). In addition, we require that

$$(B1') M_n^2 / n \to 0.$$ 

This condition ensures that the $M_n \times M_n$ matrix $n^{-1} \mathbf{X}' \mathbf{X}$ is a good estimate of its expectation (c.f. Lemma 1 below). It may be possible to weaken this to a condition like (C2) in Section 2.2 using graph theory (Mammen (1993), Lemma 1). Such an improvement is quite minor however and in practice a moderately large value $M_n$ satisfying (B1') should provide a satisfactory upper bound on the true model dimension.

**Theorem 3.** Under conditions (B1') and (B2)–(B4) in Section 2.1, $\hat{k}_0 \to_P k_0$.

4. A Simulation Study

To compare the finite-sample performance of the variable selection procedures, we carried out a simulation experiment with 1,000 trials and sample size $n = 300$ per trial.

For ordinary regression, we used the models (i) $\Gamma = \{2, 4, 5\}$, (ii) $\Gamma = \{k^2; k = 1, \ldots, 5\}$ and (iii) $\Gamma = \{2k + 7; k = 3, \ldots, 12\}$, with $M_n = 5, 30$ and 60, respectively. Each $\beta_k$, $k \in \Gamma$, was set equal to 1. The covariates were generated by a $M_n$-variate zero mean normal distribution with the $(i,j)$th entry of the covariance matrix being $2^{-|i-j|}$. The distribution of $\epsilon_i$ was standard normal. For
criterion (8), we used \( h_n(k) = kn^{0.3} \) for \( M_n = 5 \) and \( h_n(k) = kn^{0.7} \) for \( M_n = 30 \) and 60. For the latter two values of \( M_n \), the covariates were pre-ordered by their \( t \)-statistics for all four selection criteria because an all-subsets search was impractical. The simulation was coded in FORTRAN using a singular value decomposition subroutine and carried out on a SUN SPARCstation 20.

For AR model selection, we tested two cases: \( M_n = 5 \) and 30. For \( M_n = 5 \), the true model was AR(1): \( y_i = -0.3y_{i-1} + \epsilon_i \), and \( h_n(k) = kn^{0.3} \). For \( M_n = 30 \), the true model was AR(10): \( y_i = 0.2y_{i-10} + \epsilon_i \), and \( h_n(k) = kn^{0.7} \). For both cases, \( \epsilon_i \) were i.i.d. N(0,1). Initial values \( y_t, t = 1 - M_n, \ldots, 0 \), were set to zero.

Table 1. Estimated probabilities of correct model selection based on 1,000 trials; \( n = 300 \). The proposed criterion (8) is given in the last column of the table.

<table>
<thead>
<tr>
<th>( M_n )</th>
<th>True model ( \Gamma )</th>
<th>( S_p )</th>
<th>AIC</th>
<th>BIC</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>{2, 4, 5}</td>
<td>0.533</td>
<td>0.502</td>
<td>0.916</td>
<td>0.971</td>
</tr>
<tr>
<td>30</td>
<td>{k^2, k = 1, \ldots, 5}</td>
<td>0.377</td>
<td>0.365</td>
<td>0.754</td>
<td>0.929</td>
</tr>
<tr>
<td>60</td>
<td>{2k + 7, k = 3, \ldots, 12}</td>
<td>0.159</td>
<td>0.173</td>
<td>0.362</td>
<td>0.892</td>
</tr>
</tbody>
</table>

Table 1 summarizes the results for ordinary regression, which include those for the \( S_p \), the AIC (i.e., the FPE with \( \lambda = 2 \)) and the BIC criteria. The \( S_p \) and AIC criteria have fairly low probabilities of correct model selection. The BIC criterion performs better for small and moderate values of \( M_n \), but it is poor when there are many nuisance covariates (\( M_n = 60 \)). This shows the necessity for placing a heavier penalty on model complexity. The best procedure is clearly our proposed criterion (8).

Table 2. Estimated probabilities of correct AR model selection based on 1,000 trials; the proposed criterion (6) is given in the last column of the table.

<table>
<thead>
<tr>
<th>( M_n )</th>
<th>AIC</th>
<th>BIC</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.554</td>
<td>0.898</td>
<td>0.956</td>
</tr>
<tr>
<td>30</td>
<td>0.296</td>
<td>0.790</td>
<td>0.914</td>
</tr>
</tbody>
</table>

Table 2 gives the corresponding results for AR model selection. Again the proposed method is best.

Acknowledgement

This research was supported in part by U. S. Army Research Office grant DAAH04-94-G-0042.
Appendix

Proof of (3).

We only give an outline. The notation defined in Section 2 is used. Assume conditions (A1) and (A2) and that \( M_n = M \) is independent of \( n \). Then asymptotic expansions give

\[
S_p(\Theta) = \begin{cases} 
  n^{-2}e'c + n^{-2}\{2\sigma^2(|\Theta| + 2) - e'\Phi \Theta e\} + o_p(n^{-2}), & \Theta \supseteq \Gamma \\
  n^{-2}e'c + n^{-1}b'\Sigma_{\Theta}\beta + o_p(n^{-1}), & \Theta \supsetneq \Gamma.
\end{cases}
\]

Here \( \Sigma_{\Theta} = C_{\Theta} \text{diag}(0,\Psi_{\Theta})C_{\Theta}' \), where

\[
\Psi_{\Theta} = E(x_{1,\Theta}x_{1,\Theta}') - E(x_{1,\Theta}x_{1,\Theta}')[E(x_{1,\Theta}x_{1,\Theta}')]^{-1}E(x_{1,\Theta}x_{1,\Theta}')
\]

is positive definite (Seber (1984), Exercise 2.20); \( \Theta^c \) is the complement of \( \Theta \), and \( C_{\Theta} \) is a permutation matrix such that \( x_1 = C_{\Theta}(x_{1,\Theta},x_{1,\Theta}^c)' \) and \( C_{\Theta}^{-1} = C_{\Theta}' \) (see, e.g., Golub and Van Loan (1989), Section 3.4.1).

For each \( \Theta \supsetneq \Gamma \), we have \( \beta'\Sigma_{\Theta}\beta = \beta'\Psi_{\Theta}\beta_\Theta > 0 \) and \( e'\Phi_{\Theta}e = O_p(1) \). Consequently, \( \lim P(\hat{\Gamma}_{S_p} = \Theta) = 0 \) for every \( \Theta \supsetneq \Gamma \), i.e., the \( S_p \) criterion eliminates underfitting models \( \Theta \supsetneq \Gamma \).

On the other hand, if the true model is not the full model one can choose a \( \Theta \supset \Gamma \). Suppose additionally that the \( \epsilon_i \) are normally distributed. Then

\[
\lim P(\hat{\Gamma}_{S_p} = \Gamma) \leq \lim P[\sigma^{-2}e'(\Phi_{\Theta} - \Phi_{\Gamma})\epsilon \leq 2(|\Theta| - |\Gamma|)]
\]

\[
= P[\chi_{|\Theta| - |\Gamma|}^2 \leq 2(|\Theta| - |\Gamma|)] < 1.
\]

This proves the inconsistency of the \( S_p \) criterion when \( M_n \) does not depend on \( n \) and the \( \epsilon_i \) are normal. The general case when \( M_n \) may grow with \( n \) follows similarly.

Proof of Theorem 1.

Define \( \tilde{k}_0 = \arg \min_{k_0 \leq k \leq M_n} \{\text{RSS}_k + h_n(k)\hat{\sigma}^2\} \). The proof proceeds in two steps.

1. \( \tilde{k}_0 \rightarrow \rho k_0 \). For \( k \geq k_0 \), the residual sum of squares reduces to \( \text{RSS}_k = \epsilon'\epsilon - \epsilon'\Phi_k \epsilon \). Condition (B1) implies that \( E\{\epsilon'\Phi_k \epsilon/(n - M_n)\} \leq \sigma^2M_n/(n - M_n) = o(1) \) and hence

\[
\hat{\sigma}^2 = \epsilon'\epsilon n^{-1}(1 - M_n/n)^{-1} - \epsilon'\Phi_k \epsilon/(n - M_n) = \sigma^2 + o_p(1).
\]

By condition (B2) and the fact that \( (\Phi_{M_n} - \Phi_k) \) is idempotent a.s., we have for \( k > k_0 \),

\[
\text{RSS}_k + h_n(k)\hat{\sigma}^2 - \text{RSS}_{k_0} - h_n(k_0)\hat{\sigma}^2 \\
\geq \hat{\sigma}^2[h_n(k_0 + 1) - h_n(k_0)] - \epsilon'(\Phi_{M_n} - \Phi_{k_0})\epsilon.
\]
Therefore,
\[1 - P(\hat{k}_0 = k_0) \leq 1 - P\{ \min_{k > k_0} [\text{RSS}_k + h_n(k)\hat{\sigma}^2 - \text{RSS}_{k_0} - h_n(k_0)\hat{\sigma}^2] > 0 \} \]
\[\leq P[\epsilon'(\mathbf{P}_{M_n} - \mathbf{P}_{k_0})\epsilon \geq \hat{\sigma}^2 \{ h_n(k_0 + 1) - h_n(k_0) \}] \]
\[\leq P[\epsilon'(\mathbf{P}_{M_n} - \mathbf{P}_{k_0})\epsilon > \{ h_n(k_0 + 1) - h_n(k_0) \}\sigma^2/2] + P[|\hat{\sigma}^2 - \sigma^2| \geq \sigma^2/2] \]
\[\leq 2[1 - h_n(k_0)/h_n(k_0 + 1)]^{-1}\sigma^{-2}E\{\epsilon'(\mathbf{P}_{M_n} - \mathbf{P}_{k_0})\epsilon\} + o(1) \tag{15} \]
\[-2n\delta \rightarrow 0, \tag{16} \]
where (15) follows from the Markov inequality and (14), and (16) is a consequence of conditions (B2) and (B3).

2. \(\hat{k}_0 - \tilde{k}_0 = o_P(1)\). Using \(\text{RSS}_j = \epsilon'(\mathbf{I} - \mathbf{P}_j)\epsilon + \beta'\mathbf{X}'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta + 2\epsilon'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta\)
and \(\mathbf{P}_{k_0}\mathbf{X}\beta = \mathbf{X}_{k_0}\beta_{k_0} = \mathbf{X}\beta\) a.s., we get
\[P(\hat{k}_0 - \tilde{k}_0 | \neq 0) \leq \sum_{j=0}^{k_0-1} P(\hat{k}_0 = j) \]
\[\leq \sum_{j=0}^{k_0-1} P[\text{RSS}_j + h_n(j)\hat{\sigma}^2 \leq \text{RSS}_{k_0} + h_n(k_0)\hat{\sigma}^2] \]
\[\leq \sum_{j=0}^{k_0-1} P\{ -2\epsilon'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta \geq \|\mathbf{I} - \mathbf{P}_j\|\mathbf{X}\beta\|^2 - h_n(k_0)\hat{\sigma}^2 \} \]
\[\leq \sum_{j=0}^{k_0-1} P\{ -2\epsilon'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta \geq \|\mathbf{I} - \mathbf{P}_j\|\mathbf{X}\beta\|^2 - h_n(k_0)(3/2)\sigma^2 \} + k_0 P(|\hat{\sigma}^2 - \sigma^2| > \sigma^2/2) \]
\[= \sum_{j=0}^{k_0-1} P\{ -2\|\mathbf{I} - \mathbf{P}_j\|\mathbf{X}\beta\|^{-1}\epsilon'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta \geq \|\mathbf{I} - \mathbf{P}_j\|\mathbf{X}\beta\|^{-1}\} + o(1) \]
\[\leq (n\delta)^{-1/2} - (3/2)h_n(k_0)\sigma^2 n\delta^{-1/2} + \sum_{j=0}^{k_0-1} P(\|\mathbf{I} - \mathbf{P}_j\|\mathbf{X}\beta\|^2 < n\delta) + o(1) \]
\[= \sum_{j=0}^{k_0-1} R_{j1} + \sum_{j=0}^{k_0-1} R_{j2} + o(1) \]
for any constant \(\delta > 0\).
It remains to show that \( R_{j1} \) and \( R_{j2} \) both converge to zero. By the Markov inequality and condition (B4),

\[
R_{j1} \leq 4(n\delta)^{-1}[1 - 3\sigma^2 h_n(k_0)/(2n\delta)]^{-2} E\{\|\mathbf{I} - \mathbf{P}_j\mathbf{X}\beta\|^2 \geq \epsilon(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta'\mathbf{X}(\mathbf{I} - \mathbf{P}_j)\epsilon\}
\]

\[
= 4\epsilon^2(n\delta)^{-1}[1 - 3\sigma^2 h_n(k_0)/(2n\delta)]^{-2} E\{\|\mathbf{I} - \mathbf{P}_j\mathbf{X}\beta\|^2 \geq \text{tr}(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta'\mathbf{X}(\mathbf{I} - \mathbf{P}_j)\}
\]

\[
= 4\epsilon^2(n\delta)^{-1}[1 - 3\sigma^2 h_n(k_0)/(2n\delta)]^{-2} \rightarrow 0.
\]

To handle \( R_{j2} \), note that \( k_0 \) and \( j \) are independent of \( n \). Thus

\[
n^{-1}\|\mathbf{I} - \mathbf{P}_j\mathbf{X}\beta\|^2 = n^{-1}\|\mathbf{I} - \mathbf{P}_j\mathbf{X}_k\beta\|^2
\]

\[
= \beta_k\{n^{-1}\mathbf{X}_k^\prime\mathbf{X}_k - n^{-1}\mathbf{X}_k^\prime\mathbf{X}_j[n\mathbf{X}_j\mathbf{X}_j^\prime]^{-1}n^{-1}\mathbf{X}_j^\prime\mathbf{X}_k\beta\}
\]

\[
= \beta_k\Xi_j\beta_k + o_P(1),
\]

where, for \( j < k_0 \),

\[
\Xi_j = E(\mathbf{x}_k\mathbf{x}_k^\prime) - E(\mathbf{x}_k\mathbf{x}_j^\prime)[E(\mathbf{x}_j\mathbf{x}_j^\prime)]^{-1}E(\mathbf{x}_j\mathbf{x}_k^\prime).
\]

Taking \( \delta = (1/2) \min_{j<k_0} \beta_k \Xi_j, \) which is necessarily positive (see the proof of (3)), leads to \( P(\|\mathbf{I} - \mathbf{P}_j\mathbf{X}\beta\|^2 \geq n\delta) \rightarrow 1 \) for \( j < k_0 \). This implies \( R_{j2} = o(1) \) and hence completes the proof.

**Proof of Theorem 2.**

It suffices to prove (9). Let \( \xi \) denote the maximum eigenvalue of \( \mathbf{A}'\mathbf{A} \) and let \( \|\mathbf{A}\| = \xi^{1/2} \) denote the matrix 2-norm of \( \mathbf{A} \). Further, let \( \mathbf{e}_i \) be the \( M_n \)-dimensional unit vector with its \( i \)th component equal to 1. By Lemma 1 of Mammen (1993) and condition (C2), \( \|\mathbf{B}\| = o_P(1), \) where

\[
\mathbf{B} = [E(\mathbf{x}_1\mathbf{x}_1^\prime)]^{-1/2}(n^{-1}\mathbf{X}'\mathbf{X})[E(\mathbf{x}_1\mathbf{x}_1^\prime)]^{-1/2} - \mathbf{I}.
\]

Therefore by Theorem 10.3.1 of Campbell and Meyer (1979), with probability tending to one, \( \mathbf{B} + \mathbf{I} \) and consequently \( \mathbf{X}'\mathbf{X} \) are invertible. Without loss of generality we therefore assume that \( (\mathbf{X}'\mathbf{X})^{-1} \) exists. Write \( T_1 = \hat{\beta}^{-1}\hat{\beta}_i[\mathbf{e}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i]^{-1/2} \) and \( \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon. \)

If \( i \not\in \Gamma \), then \( \hat{\beta}_i = 0 \) and \( \hat{\beta}_i = \mathbf{e}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon. \) By first conditioning on \( \mathbf{X} \), we see that \( E(\hat{\sigma}T_1) = 0 \) and

\[
E(\hat{\sigma}^2T_1^2) = E[\mathbf{e}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i]^{-1}E[\epsilon'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i\mathbf{e}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon|\mathbf{X}]
\]

\[
= \sigma^2.
\]
Therefore by condition (C2),
\[ P\left[ \max_{i \in \Gamma} \hat{\sigma}[T_i] \geq n^{\frac{1}{2(1+\eta)}} \right] \leq \sum_{i \in \Gamma} P(\hat{\sigma}[T_i] \geq n^{\frac{1}{2(1+\eta)}}) \leq \sum_{i \in \Gamma} E(\hat{\sigma}^2 T_i) n^{-1/(1+\eta)} \leq \sigma^2 M n^{-1/(1+\eta)} = o(1). \] (18)

On the other hand, if \( i \in \Gamma \),
\[
\hat{\sigma}T_i = \beta_i \{ e_i'(X'X)^{-1}e_i \}^{-1/2} + \epsilon_i'X'X^{-1}X'\{ e_i'(X'X)^{-1}e_i \}^{-1/2} = \beta_i \{ e_i'(X'X)^{-1}e_i \}^{-1/2} + O_P(1),
\] (19)
by (17). Since \( \|B\| = o_P(1) \), we have
\[
e_i' (n^{-1}X'X)^{-1}e_i = e_i'[E(x_1x_1') - |E(x_1x_1') - n^{-1}X'X|]^{-1}e_i
= e_i'[E(x_1x_1')]^{-1/2}(I + B)^{-1}[E(x_1x_1')]^{-1/2}e_i
= e_i'[E(x_1x_1')]^{-1/2}(I + \sum_{j \geq 1}(-1)^jB^j)[E(x_1x_1')]^{-1/2}e_i
= e_i'[E(x_1x_1')]^{-1}e_i + R_i,
\]
where (c.f. Mammen (1993) or Golub and Van Loan (1989), Section 2.3)
\[
|R_i| = \left| e_i'[E(x_1x_1')]^{-1/2} \sum_{j \geq 1}(-1)^jB^j[E(x_1x_1')]^{-1/2}e_i \right|
\leq \left\| [E(x_1x_1')]^{-1/2} \right\|^2 \cdot \left\| \sum_{j \geq 1} B^j \right\|
\leq \left\| [E(x_1x_1')]^{-1/2} \right\|^2 \cdot \|B\|(1 - \|B\|)^{-1} = o_P(1)
\]
by (C1). Equality (19) implies that, for \( i \in \Gamma \),
\[
\hat{\sigma}n^{-1/2}T_i = \beta_i \{ e_i'[E(x_1x_1')]^{-1}e_i + o_P(1) \}^{-1/2} + o_P(1)
= \beta_i \{ e_i'[E(x_1x_1')]^{-1}e_i \}^{-1/2} + o_P(1).
\]
Note that the first term on the right side of the last equality is bounded away from zero because by condition (C1), \( e_i'[E(x_1x_1')]^{-1}e_i \leq \kappa^{-1} < \infty \). Since \( \Gamma \) is independent of \( n \), it follows that
\[
P(\min_{i \in \Gamma} \hat{\sigma}[T_i] \geq n^{\frac{1}{2(1+\eta)}}) \rightarrow 1, \quad \eta > 0.
\]
This and (18) yield (9).

**Proof of Theorem 3.**

The method used to prove Theorem 1 still works here with some technical modifications to accommodate dependency among observations. We follow the two steps there with the same definition of \( \bar{k}_0 \).
1. $\hat{k}_0 \rightarrow P k_0$. As in the proof of Theorem 1, (c.f. the inequality before (15)), we have

$$1 - P(\hat{k}_0 = k_0) \leq P[\epsilon' P_{M_n} \epsilon > \{h_n(k_0 + 1) - h_n(k_0)\} \sigma^2/2] + P(\{\sigma^2 - \sigma^2\} \geq \sigma^2/2),$$

which converges to zero if we can show that (c.f. (14))

$$\epsilon' P_{M_n} \epsilon = O_p(M_n). \quad (20)$$

Write the autocovariance matrix $E(\epsilon_1 \epsilon'_1) = \Sigma_{M_n} = \gamma_0(\sigma_{ij})_{1 \leq i, j \leq M_n}$ where $\gamma_0 = \text{var}(y_1)$ and $\sigma_{ij} = \rho_{|i-j|}$ is the autocorrelation defined in (12). It is known that $\Sigma_{M_n}$ is nonsingular for all $1 \leq M_n \leq n$ (Hannan (1973)). Let $B = \Sigma_{M_n}^{-1/2}(n^{-1}XX')\Sigma_{M_n}^{-1/2} - I$. From Lemma 1 below, $\|B\| = o_P(1)$. Now (20) becomes

$$\epsilon' P_{M_n} \epsilon = n^{-1}(\Sigma_{M_n}^{-1/2}X'\epsilon)'(I + B)^{-1}(\Sigma_{M_n}^{-1/2}X'\epsilon)$$

$$= n^{-1}(\Sigma_{M_n}^{-1/2}X'\epsilon)'(I + \sum_{j \geq 1}(-B)^{j})(\Sigma_{M_n}^{-1/2}X'\epsilon)$$

$$= n^{-1}\|\Sigma_{M_n}^{-1/2}X'\epsilon\|^2 + R,$$

where (c.f. the proof of Theorem 2)

$$|R| \leq n^{-1}\|\Sigma_{M_n}^{-1/2}X'\epsilon\|^2 \|B\|(1 - \|B\|)^{-1} = n^{-1}\|\Sigma_{M_n}^{-1/2}X'\epsilon\|^2 \cdot o_P(1).$$

It thus suffices to establish that

$$n^{-1}\|\Sigma_{M_n}^{-1/2}X'\epsilon\|^2 = O_P(M_n). \quad (21)$$

Recall that $\epsilon_i$ is independent of $\{y_{i-k}, k \geq 1\}$ and hence of $x_i' = (y_{i-1}, \ldots, y_{i-M_n})$ for a stationary AR$(k_0)$ process, and that each $y_j$ can be written as a linear combination of $\{\epsilon_t, t \leq j\}$ (Box and Jenkins (1976)). Therefore

$$E(\epsilon' \epsilon'X) = E\{\sum_{i=1}^n x_i \epsilon_i \sum_{j=1}^n \epsilon_j' x_i'\} = \sum_{i=1}^n E(\epsilon_i' x_i x_i') = n\sigma^2 \Sigma_{M_n},$$

where each $E(\epsilon_i \epsilon_i \epsilon_j x_j') = 0$ for $i \neq j$ by first conditioning on $\{\epsilon_t, t \leq \max(i,j) - 1\}$. The expectation of the left hand side of (21) is thus

$$n^{-1}E\{\epsilon' X \Sigma_{M_n}^{-1} X' \epsilon\} = n^{-1} \text{tr}\{\Sigma_{M_n}^{-1} E[XX' \epsilon' \epsilon]\} = \sigma^2 M_n.$$

This proves (21).
2. \( \hat{k}_0 - \tilde{k}_0 = o_P(1) \). As in Step 2 of the proof of Theorem 1,

\[
P(\{\hat{k}_0 - \tilde{k}_0 \neq 0\}) \leq \sum_{j=0}^{k_0-1} P\{n^{-1}\text{RSS}_j - n^{-1}\text{RSS}_{k_0} \leq n^{-1}[h_n(k_0) - h_n(j)]\hat{\sigma}^2\}.
\]

(22)

It follows from Potscher (1989), proof of Lemma 3.3 that for each \( j < k_0 \),

\[\lim \inf_n \{n^{-1}\text{RSS}_j - n^{-1}\text{RSS}_{k_0}\} > 0, \text{ a.s.}\]

Condition (B4) and the fact that \( \hat{\sigma}^2 \to P \sigma^2 \) imply that (22) converges to zero.

**Lemma 1.** Suppose that the conditions of Theorem 3 hold. Then

1. The minimum eigenvalue \( \lambda_{M_n} \) of \( \Sigma_{M_n} \) satisfies \( \lambda_{M_n} \geq \lambda > 0 \) for some constant \( \lambda \) independent of \( n \).
2. \( \|B\| = o_P(1) \).

**Proof.** The proof makes use of Toeplitz forms in Grenander and Szego (1984), Chapters 5 and 10 and is available from the first author.

**References**


Department of Mathematics and Statistics, Utah State University, Logan, UT 84322-3900, U.S.A.

Department of Statistics, University of Wisconsin, Madison, WI 53706. U.S.A.

(Received December 1994; accepted October 1996)