TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 347, Number 2, February 1995

ASSESSING PREDICTION ERROR IN AUTOREGRESSIVE MODELS

PING ZHANG AND PAUL SHAMAN

ABSTRACT. Assessing prediction error is a problem which arises in time series analysis. The distinction between the conditional prediction error e and the unconditional prediction error $E(e)$ has not received much attention in the literature. Although one can argue that the conditional version is more practical, we show in this article that assessing e is nearly impossible. In particular, we use the correlation coefficient $\operatorname{corr}(\hat{e}, e)$, where $\hat{e}$ is an estimate of e, as a measure of performance and show that $\lim_{T\to\infty} \sqrt{T}\operatorname{corr}(\hat{e}, e) = C$ where T is the sample size and $C > 0$ is some constant. Furthermore, the value of C is large only when the process is extremely non-Gaussian or nearly nonstationary.

1. INTRODUCTION

One of the goals of time series modelling is to forecast future values. The idea of using prediction error as a measure of performance is very popular in statistics. One notable example is the area known as model selection. Many model selection criteria are derived with the aim of achieving minimal prediction error. These include the AIC criterion [2] and the final prediction error (FPE) criterion [1]. See Linhart and Zucchini [7] for further discussion on this topic. In practice, since the true prediction error is unknown, one uses estimated prediction error. Here in this article, the true prediction error is referred to as the conditional prediction error, which depends on both the unknown distribution and the past observations. Johnstone [6] points out that many natural estimates of prediction error are not admissible. Another observation made by Shibata [10] shows that many prediction error-based model selection criteria are inconsistent. This raises the important question of whether prediction error is estimable.

Consider the pth-order autoregressive process with mean zero,

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t, \qquad t = 0, \pm 1, \pm 2, \ldots, \tag{1}$$

where $\{\varepsilon_t\}$ is a $(0, \sigma^2)$ white noise process. Let

$$1 - \alpha_1 z - \cdots - \alpha_p z^p = (1 - \phi_1 z) \cdots (1 - \phi_p z),$$

Received by the editors April 16, 1993; originally communicated to the *Proceedings of the AMS* by Wei Y. Loh.

1991 *Mathematics Subject Classification.* Primary 62M10, 62M20; Secondary 62H20.

Key words and phrases. Conditional prediction error, correlation, cumulant, higher-order spectrum, non-Gaussian model.

Research of the first author was partially supported by the Research Foundation of the University of Pennsylvania and NSA Grant No. MDA904-93-H-3014.

0002-9947/95 $1.00 + $.25 per page

where $\phi_1, \ldots, \phi_p$ lie strictly within the unit circle. Suppose that we have observed $(X_1, \ldots, X_T)$ and would like to predict the value of X_{T+1}. Let $\mathcal{F}_t$ be the σ-field generated by $\{X_s, s \leq t\}$, and denote the predicted value of X_{T+1} by $\hat{X}_{T+1}$. Then the conditional prediction error can be written as

$$e = E\{(X_{T+1} - \hat{X}_{T+1})^2 | \mathcal{F}_T\}. \tag{2}$$

Let $\hat{\Gamma}$ be the $p \times p$ matrix with the (i, j)th element equal to $\sum_{t=p+1}^{T} X_{t-i}X_{t-j}$, and let $\hat{\gamma}$ be the $p \times 1$ vector with the ith element equal to $\sum_{t=p+1}^{T} X_t X_{t-i}$. Then the least squares predictor has the form

$$\hat{X}_{T+1} = \hat{\alpha}_1 X_T + \cdots + \hat{\alpha}_p X_{T-p+1} = \hat{\alpha}' \tilde{X}_T,$$

where $\tilde{X}_T = (X_T, X_{T-1}, \ldots, X_{T-p+1})'$ and $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)'$ satisfies $\hat{\Gamma}\hat{\alpha} = \hat{\gamma}$. Notice that $X_{T+1} = \alpha' \tilde{X}_T + \varepsilon_{T+1}$ and that (2) can be expressed as

$$e = \sigma^2 + \{(\hat{\alpha} - \alpha)' \tilde{X}_T\}^2.$$

In order to derive an estimate of e, we appeal to a result of Fuller and Hasza [5, Theorem 2.1]. If $\{\varepsilon_t\}$ is Gaussian, it has been shown that

$$(\hat{\alpha} - \alpha)(\hat{\alpha} - \alpha)' = T^{-1}\sigma^2\Gamma^{-1} + O_p(T^{-3/2}),$$

where Γ is the $p \times p$ matrix with the (i, j)th element equal to $E(X_{t-i}X_{t-j})$. Consequently, for Gaussian $AR(p)$ processes, we have

$$e = \sigma^2\{1 + T^{-1}\tilde{X}_T'\Gamma^{-1}\tilde{X}_T\} + O_p(T^{-3/2}). \tag{3}$$

Under the Gaussian assumption, one can further show that

$$E(e) = \sigma^2(1 + pT^{-1}) + O(T^{-3/2}).$$

Hence we use

$$\hat{e} = \hat{\sigma}^2(1 + pT^{-1}) \tag{4}$$

to estimate e, where $\hat{\sigma}^2 = (T - 2p)^{-1}\sum_{t=p+1}^{T}(X_t - \hat{\alpha}'\tilde{X}_{t-1})^2$. In the model selection context, the FPE criterion of Akaike [1] minimizes (4) as p varies.

The question arises of how to measure the performance of $\hat{e}$ as an estimate of e. A trivial observation is that both $\hat{e}$ and e tend to σ^2 as $T \to \infty$. Shaman [8] derives the asymptotic expectation and variance of $\hat{e}$, and the results show in general that $\hat{e}$ and e have the same expectation, but not the same variance, to terms of order T^{-1}. Another approach is to look at the distance between $\hat{e}$ and e. Johnstone [6] shows that natural estimates such as (4) can be inadmissible for squared distance $E(\hat{e} - e)^2$. In this article, since both $\hat{e}$ and e are random variables, we propose using the correlation coefficient $\operatorname{corr}(\hat{e}, e)$ as a measure of performance.

The layout of this article is as follows. In §2, we introduce some preliminary results concerning the frequency domain properties of a linear stationary time series. These will make later calculations much more transparent. Our major conclusions, however, are stated in time domain language. Sections 3 and 4 deal with Gaussian and non-Gaussian autoregressive processes, respectively. In both cases, we show that $\operatorname{corr}(\hat{e}, e) = O(T^{-1/2})$. As an implication, $\hat{e}$ and e are asymptotically uncorrelated. Section 5 offers some explanations and remarks.

2. Cumulants and spectra

This section introduces some results concerning the frequency domain properties of a linear stationary time series. This effort is well worth it because the proofs of later results, although stated in time domain language, become much more transparent when the frequency domain representation is used.

Consider a random vector $X = (X_1, X_2, \ldots, X_m)$ with generating function

$$g(\theta) = E\{\exp(\theta \circ X)\},$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ and $\theta \circ X = \sum \theta_i X_i$. Suppose that $\log g(\theta)$ admits a Taylor expansion,

$$\log g(\theta) = \sum_r \kappa_r \{\theta^r / r!\}.$$

The coefficients κ_r are called the cumulants of X. Here $r = (r_1, \ldots, r_m)$ is a vector of nonnegative integers, $\theta^r = \prod_{i=1}^m \theta_i^{r_i}$, and $r! = \prod_{i=1}^m r_i!$. It is well known that the set of all cumulants is equivalent to the set of all product moments in the sense that any product moment can be expressed as a function of cumulants and vice versa. Define $\operatorname{cum}(X_1, \ldots, X_m) = \kappa_{(1,\ldots,1)}$. Suppose that X_1, X_2, X_3, X_4 all have mean zero. Then

$$E(X_1) = \operatorname{cum}(X_1), \qquad E(X_1X_2) = \operatorname{cum}(X_1, X_2),$$
$$E(X_1X_2X_3) = \operatorname{cum}(X_1, X_2, X_3),$$

and

$$\begin{aligned} E(X_1X_2X_3X_4) = {} & \operatorname{cum}(X_1, X_2, X_3, X_4) + \operatorname{cum}(X_1, X_2)\operatorname{cum}(X_3, X_4) \\ & + \operatorname{cum}(X_1, X_3)\operatorname{cum}(X_2, X_4) + \operatorname{cum}(X_1, X_4)\operatorname{cum}(X_2, X_3). \end{aligned}$$

Historically, it is folklore that almost all meaningful statistical parameters are nice functions of the cumulants. The notion of multiple cumulants has proved to be very helpful in getting insight in the context of analysis of variance [11]. Higher-order cumulants are also used extensively in spectral analysis of time series. A brief history of this development is given by Brillinger [4].

Suppose that $\{X_t\}$ is a stationary process with mean zero and spectral density $f(\omega)$. It is well known that X_t has the spectral representation

$$X_t = \int_{-\pi}^{\pi} e^{i\omega t}\, dZ(\omega),$$

where $Z(\omega)$ is a complex-valued orthogonal process satisfying $E[dZ(\omega)] = 0$ and $\operatorname{cum}(dZ(\omega), dZ(\omega')) = \delta(\omega+\omega')f(\omega)\, d\omega\, d\omega'$, where δ is the Dirac delta

function. As a generalization, higher-order spectral densities are defined by

$$\operatorname{cum}\{dZ(\lambda_1), \ldots, dZ(\lambda_k)\} = \eta(\lambda_1 + \cdots + \lambda_k) f_k(\lambda_1, \ldots, \lambda_{k-1})\, d\lambda_1 \cdots d\lambda_k,$$

where $\eta(\lambda)$ is the Dirac comb defined by

$$\eta(\lambda) = \sum_{l=-\infty}^{\infty} \delta(\lambda + 2\pi l).$$

For Gaussian stationary processes, $f_k(\lambda_1, \ldots, \lambda_{k-1}) \equiv 0$ if $k > 2$. One use of higher-order spectral densities is therefore to study nonlinear or non-Gaussian processes. The following result holds.

Theorem 1. *Suppose that* $X_t = \int \exp(i\omega t)\, dZ(\omega)$ *is a stationary process with mean zero and spectral density* $f(\lambda)$. *Let* $f_4(\omega, \lambda, \mu)$ *be the fourth-order spectral density of* $\{X_t\}$. *Let* $a(\omega)$ *and* $b(\omega)$ *be complex-valued functions. Then*

$$\operatorname{cov}\left\{\left(\int a(\omega)\, dZ(\omega)\right)^2, \left(\int b(\omega)\, dZ(\omega)\right)^2\right\} = K(f_4) + 2\left\{\int a(\omega) b(-\omega) f(\omega)\, d\omega\right\}^2,$$

provided that the integral on the right-hand side exists. Here $K(f_4)$ *is a functional of* f_4 *given by*

$$\begin{aligned} K(f_4) = &\int_{R_1} a(\omega)a(\lambda)b(\mu)b(-\omega-\lambda-\mu) f_4(\omega, \lambda, \mu)\, d\omega\, d\lambda\, d\mu \\ &+ \int_{R_2} a(\omega)a(\lambda)b(\mu)b(2\pi-\omega-\lambda-\mu) f_4(\omega, \lambda, \mu)\, d\omega\, d\lambda\, d\mu \\ &+ \int_{R_3} a(\omega)a(\lambda)b(\mu)b(-2\pi-\omega-\lambda-\mu) f_4(\omega, \lambda, \mu)\, d\omega\, d\lambda\, d\mu, \end{aligned}$$

where

$$R_1 = \{(\omega, \lambda, \mu): -\pi \le \omega, \lambda, \mu \le \pi, -\pi \le \omega + \lambda + \mu \le \pi\},$$

$$R_2 = \{(\omega, \lambda, \mu): -\pi \le \omega, \lambda, \mu \le \pi, \pi \le \omega + \lambda + \mu \le 3\pi\},$$

and

$$R_3 = \{(\omega, \lambda, \mu): -\pi \le \omega, \lambda, \mu \le \pi, -3\pi \le \omega + \lambda + \mu \le -\pi\}.$$

Proof. Expressing product moments in terms of cumulants, we obtain the conclusion from direct calculation. To be specific, we have

$$
\begin{aligned}
E\left\{\left(\int a(\omega)\,dZ(\omega)\right)^2\cdot\left(\int b(\omega)\,dZ(\omega)\right)^2\right\}
&= \int a(\omega)a(\lambda)b(\mu)b(\nu)E\{dZ(\omega)\,dZ(\lambda)\,dZ(\mu)\,dZ(\nu)\}\\
&= \int a(\omega)a(\lambda)b(\mu)b(\nu)\{\operatorname{cum}(dZ(\omega),dZ(\lambda),dZ(\mu),dZ(\nu))\\
&\quad+\operatorname{cum}(dZ(\omega),dZ(\mu))\operatorname{cum}(dZ(\lambda),dZ(\nu))\\
&\quad+\operatorname{cum}(dZ(\omega),dZ(\nu)))\operatorname{cum}(dZ(\lambda),dZ(\mu))\\
&\qquad\qquad+\operatorname{cum}(dZ(\omega),dZ(\lambda))\operatorname{cum}(dZ(\mu),dZ(\nu))\}\\
&= K(f_4)+2\int a(\omega)a(\lambda)b(-\omega)b(-\lambda)f(\omega)f(\lambda)\,d\omega\,d\lambda\\
&\quad+\int a(\omega)a(-\omega)b(\lambda)b(-\lambda)f(\omega)f(\lambda)\,d\omega\,d\lambda\\
&= K(f_4)+2\left\{\int a(\omega)b(-\omega)f(\omega)\,d\omega\right\}^2\\
&\quad+\left\{\int a(\omega)a(-\omega)f(\omega)d\omega\right\}\cdot\left\{\int b(\omega)b(-\omega)f(\omega)\,d\omega\right\}\\
&= K(f_4)+2\left\{\int a(\omega)b(-\omega)f(\omega)\,d\omega\right\}^2\\
&\quad+E\left(\int a(\omega)\,dZ(\omega)\right)^2\cdot E\left(\int b(\omega)\,dZ(\omega)\right)^2.
\end{aligned}
$$

Moving the last term to the left-hand side, we obtain the conclusion. To establish the expression for $K(f_4)$, notice that

$$
\begin{aligned}
&\operatorname{cum}(dZ(\omega),dZ(\lambda),dZ(\mu),dZ(\nu))\\
&\quad=\eta(\omega+\lambda+\mu+\nu)f_4(\omega,\lambda,\mu)\,d\omega\,d\lambda\,d\mu\,d\nu,
\end{aligned}
$$

where $\eta(\cdot)$ equals zero unless its argument is a multiple of 2π. In our case, since ω,λ,μ,ν are within $-\pi$ and π, the only nonzero terms left after integrating both sides of the above equation correspond to the cases where $\omega+\lambda+\mu+\nu=0$, 2π, and -2π. □

We shall also need the following result.

Theorem 2. *Suppose that* $\{\varepsilon_t\}$ *is a sequence of iid* $(0,\sigma^2)$ *random variables with fourth cumulant* κ_4. *Let* $f_4(\omega,\lambda,\mu)$ *be the fourth-order spectral density of* $\{\varepsilon_t\}$. *Then*

$$
\int_{R_1\cup R_2\cup R_3} e^{i\omega r}e^{i\lambda s}e^{i\mu t}f_4(\omega,\lambda,\mu)\,d\omega\,d\lambda\,d\mu=\kappa_4 I\{(r,s,t)=(0,0,0)\},
$$

where $I(\cdot)$ *is an indicator function.*

Proof. By definition,

$$
\int_{R_1\cup R_2\cup R_3} e^{i\omega r}e^{i\lambda s}e^{i\mu t}f_4(\omega,\lambda,\mu)\,d\omega\,d\lambda\,d\mu=\operatorname{cum}(\varepsilon_0,\varepsilon_r,\varepsilon_s,\varepsilon_t).
$$

Since $\{\varepsilon_t\}$ is an iid sequence, the right-hand side is nonzero if and only if $r = s = t = 0$. Obviously, $\operatorname{cum}(\varepsilon_0, \varepsilon_0, \varepsilon_0, \varepsilon_0) = \kappa_4$. □

3. Gaussian autoregressive models

Suppose that $\{\varepsilon_t\}$ is a white noise process having spectral representation $\varepsilon_t = \int e^{i\omega t}\, dZ_0(\omega)$. Then $Z_0(\omega)$ is an orthogonal process with $E|dZ_0(\omega)|^2 = \sigma^2(2\pi)^{-1}\, d\omega$. Let X_t be the linear stationary process

$$X_t = \sum_{k=0}^{\infty} \xi_k \varepsilon_{t-k}, \tag{5}$$

where $\xi_0 = 1$. Then the spectral representation for X_t is

$$X_t = \int e^{i\omega t} A(e^{-i\omega})\, dZ_0(\omega),$$

where $A(e^{-i\omega}) = \sum_{k=0}^{\infty} \xi_k e^{-ik\omega}$ is the transfer function. Suppose that $h(\omega)$ is the spectral density of $\{X_t\}$. Then

$$h(\omega) = \sigma^2(2\pi)^{-1}|A(e^{-i\omega})|^2.$$

For illustration, let us first consider the simple case of the $AR(1)$ model,

$$X_t = \alpha X_{t-1} + \varepsilon_t, \qquad t = 0, \pm 1, \ldots.$$

Now equation (3) becomes

$$e = \sigma^2 + T^{-1}X_T^2(1 - \alpha^2) + O_p(T^{-3/2}).$$

It is easy to see that $\hat{e} = (T-2)^{-1}(1 + T^{-1})(\sum_{t=2}^{T} \varepsilon_t^2) \cdot (1 + o_p(1))$. Thus

$$\lim_{T\to\infty} \sqrt{T}\operatorname{corr}(\hat{e}, e) = \lim_{T\to\infty} \sqrt{T}\operatorname{corr}\left(X_T^2, \sum_{t=2}^{T} \varepsilon_t^2\right).$$

To find this limit, we appeal to Theorem 1 with $a(\omega) = e^{iT\omega}A(e^{-i\omega})$, $b(\omega) = e^{i\omega t}$, and $f(\omega) = \sigma^2/(2\pi)$. For Gaussian processes, $K(f_4) = 0$. Hence

$$\operatorname{cov}(X_T^2, \varepsilon_t^2) = 2\left\{\frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi} e^{i(T-t)\omega}A(e^{-i\omega})\, d\omega\right\}^2 = 2\sigma^4\alpha^{2(T-t)},$$

where the last equality is due to the fact that $A(e^{-i\omega}) = (1 - \alpha e^{-i\omega})^{-1}$. A similar argument yields that

$$\operatorname{var}(X_T^2) = 2\left\{\frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi} |A(e^{i\omega})|^2\, d\omega\right\}^2 = 2\sigma^4(1 - \alpha^2)^{-2}.$$

Finally, since $\operatorname{var}(\sum_{t=2}^{T} \varepsilon_t^2) = 2(T-1)\sigma^4$, it follows that

$$\lim_{T\to\infty} \sqrt{T}\operatorname{corr}(\hat{e}, e) = 1.$$

The previous argument applied to the case of general $AR(p)$ models entails the following result.

Theorem 3. *Suppose that* $\{X_t\}$ *is a stationary Gaussian* $AR(p)$ *model. Let* e *and* $\hat{e}$ *be as defined earlier. Let* $\operatorname{corr}(\hat{e}, e)$ *be the correlation coefficient between* $\hat{e}$ *and* e. *Then*

$$\lim_{T\to\infty} \sqrt{T}\operatorname{corr}(\hat{e}, e) = \sqrt{p}.$$

Proof. It is easy to show that

$$\hat{e} = (T-2p)^{-1}(1+pT^{-1})\left(\sum_{t=p+1}^{T} \varepsilon_t^2\right)\cdot(1+o_p(1)).$$

Thus by (3), we only need to show that

$$\lim_{T\to\infty} \sqrt{T}\operatorname{corr}\left(\tilde{X}_T'\Gamma^{-1}\tilde{X}_T, \sum_{t=p+1}^{T} \varepsilon_t^2\right) = \sqrt{p}.$$

Define $e(\omega) = (e^{iT\omega}, e^{i(T-1)\omega}, \ldots, e^{i(T-p+1)\omega})'$. Then

$$\tilde{X}_T'\Gamma^{-1}\tilde{X}_T = \|\Gamma^{-1/2}\tilde{X}_T\|^2 = \left\|\Gamma^{-1/2}\int e(\omega)A(e^{-i\omega})\,dZ_0(\omega)\right\|^2.$$

Also, one can write

$$\sum_{t=p+1}^{T} \varepsilon_t^2 = \sum_{t=p+1}^{T}\left\{\int e^{it\omega}\,dZ_0(\omega)\right\}^2.$$

Using Theorem 1, it is easy to show that

$$\operatorname{var}(\tilde{X}_T'\Gamma^{-1}\tilde{X}_T) = 2tr\left\{\Gamma^{-1/2}\frac{\sigma^2}{2\pi}\int e(\omega)e'(-\omega)|A(e^{-i\omega})|^2\,d\omega\Gamma^{-1/2}\right\}^2 = 2p,$$

where the last equality is due to the fact that

$$\int e(\omega)e'(-\omega)h(\omega)\,d\omega = \Gamma.$$

Next, Theorem 1 also implies that

$$\begin{aligned}\operatorname{cov}(\tilde{X}_T'\Gamma^{-1}\tilde{X}_T, \varepsilon_t^2) &= 2\left\|\frac{\sigma^2}{2\pi}\int \Gamma^{-1/2}e(\omega)A(e^{-i\omega})e^{-it\omega}\,d\omega\right\|^2 \\ &= 2\sigma^4\xi'(T-t)\Gamma^{-1}\xi(T-t),\end{aligned}$$

where

$$\xi(t) = (\xi_t, \ldots, \xi_{t-p+1})',$$

and the ξ_t's are as in (5) with $\xi_k = 0$ if $k < 0$. This implies that

$$\begin{aligned}\operatorname{cov}\left(\tilde{X}_T'\Gamma^{-1}\tilde{X}_T, \sum_{t=p+1}^{T} \varepsilon_t^2\right) &= 2\sigma^4 tr\left\{\Gamma^{-1}\sum_{t=p+1}^{T} \xi(T-t)\xi'(T-t)\right\} \\ &= 2\sigma^2 p(1+o(1)).\end{aligned}$$

Finally, since $\operatorname{var}(\sum_{t=p+1}^{T} \varepsilon_t^2) = 2(T-p)\sigma^4$, we get

$$\operatorname{corr}\left(\widetilde{X}_T'\Gamma^{-1}\widetilde{X}_T, \sum_{t=p+1}^{T} \varepsilon_t^2\right) = \frac{\operatorname{cov}(\widetilde{X}_T'\Gamma^{-1}\widetilde{X}_T, \sum_{t=p+1}^{T} \varepsilon_t^2)}{\{\operatorname{var}(\widetilde{X}_T'\Gamma^{-1}\widetilde{X}_T)\}^{1/2} \cdot \{\operatorname{var}(\sum_{t=p+1}^{T} \varepsilon_t^2)\}^{1/2}}$$
$$= \frac{2\sigma^2 p}{\sqrt{2p} \cdot \sqrt{2T\sigma^4}} \cdot (1 + o(1)) = \sqrt{p/T} \cdot (1 + o(1)).$$

The conclusion follows immediately. □

4. Non-Gaussian autoregressive models

The key to the exposition in previous sections is equation (3), which holds when the innovation term ε_t in (1) is Gaussian $(0, \sigma^2)$. To establish results for general non-Gaussian autoregressive models, we slightly modify the definition of the prediction rule. Let $\delta_T > 0$ be a constant that converges to zero as $T \to \infty$. Let $\widetilde{X}_T$, $\widehat{\Gamma}$, and $\hat{\gamma}$ be as given in §1. Throughout this section, we define the conditional prediction error by (2) where

$$\widehat{X}_{T+1} = \hat{\gamma}'(\widehat{\Gamma} + \delta_T I)^{-1}\widetilde{X}_T.$$

Here I is the identity matrix. Notice that when $\delta_T = 0$, the above prediction rule reduces to the least squares rule. For the modified prediction error, the following result holds.

Lemma 1. *Suppose that ε_t in (1) is a sequence of $iid(0, \sigma^2)$ random variables following a symmetric distribution and $E\exp(s|\varepsilon_t|) < \infty$ for some $s > 0$. Suppose that there exist positive constants c_1 and c_2 such that $c_1 T^{-k} < \delta_T < c_2 T^{-1/2}$ for some $k > 1/2$. Then equation (3) holds.*

Proof. Without losing generality, we only sketch a proof for the $AR(1)$ case. General $AR(p)$ models can be dealt with in a similar fashion. Let $\tilde{\alpha} = (\widehat{\Gamma} + \delta_T I)^{-1}\hat{\gamma}$. Then all we need to show is

$$(\tilde{\alpha} - \alpha)(\tilde{\alpha} - \alpha)' = T^{-1}\sigma^2\Gamma^{-1} + O_p(T^{-3/2}).$$

Now if $p = 1$, then one can easily verify that

$$\Gamma = \sigma_X^2 = EX_1^2 = \sigma^2(1 - \alpha^2)^{-1}.$$

Moreover, we can write

$$\tilde{\alpha} - \alpha = -\alpha\delta_T\left(\sum_{t=2}^{T} X_{t-1}^2 + \delta_T\right)^{-1} + \left(\sum_{t=2}^{T} X_{t-1}^2 + \delta_T\right)^{-1}\sum_{t=2}^{T} \varepsilon_t X_{t-1} = \xi_T + \eta_T.$$

Define

$$A = \left\{\left|T^{-1}\sum_{t=2}^{T} X_{t-1}^2 - \sigma_X^2\right| < \sigma_X^2/2\right\}.$$

For samples in A, we have $\sum_{t=2}^{T} X_{t-1}^2 > T\sigma_X^2/2$. Hence $|\xi_T|I_A < C\delta_T T^{-1}$ for some constant C. Next, since $|\xi_T| < |\alpha|$, we have for any integer l that

$$E|\xi_T|^l I_{\overline{A}} \le |\alpha|^l P(\overline{A}) = |\alpha|^l P\left\{\left|T^{-1}\sum_{t=2}^{T} X_{t-1}^2 - \sigma_X^2\right| \ge \frac{\sigma_X^2}{2}\right\} = O(T^{-3l/2}).$$

Under the assumptions, one can actually obtain an exponential bound for $P(\overline{A})$. But the above bound suffices for later argument. See also [9]. Let $L_l(C)$ denote the set of random variables whose lth moment is bounded by C. The above arguments imply that $T^{3/2}\xi_T \in L_l(C)$. Likewise, it is easy to show that

$$T^{3/2}\left\{\eta_T - (\sigma_X^2 T)^{-1}\sum_{t=2}^{T}\varepsilon_t X_{t-1}\right\} \in L_l(C).$$

Consequently, we can express $\tilde{\alpha} - \alpha$ as

$$\tilde{\alpha} - \alpha = \hat{\theta}_T T^{-3/2} + (\sigma_X^2 T)^{-1}\sum_{t=2}^{T}\varepsilon_t X_{t-1},$$

where $\hat{\theta}_T \in L_l(C)$. Straightforward calculation yields that

$$E(\tilde{\alpha} - \alpha)^2 = T^{-1}\sigma^2\sigma_X^{-2} + O(T^{-2})$$

and

$$\operatorname{var}\{(\tilde{\alpha} - \alpha)^2\} = O(T^{-3}).$$

The conclusion follows from the Cauchy-Schwarz inequality. □

Following Lemma 1, the correlation coefficient between e and $\hat{e}$ can be approximated by

$$\operatorname{corr}\left(\tilde{X}_T'\Gamma^{-1}\tilde{X}_T, \sum_{t=p+1}^{T}\varepsilon_t^2\right).$$

The latter expression can then be analyzed using higher-order spectral analysis. The following is the main result of this section.

Theorem 4. *Under the assumptions of Lemma* 1, *let* κ_4 *be the fourth cumulant of* ε_t. *Then*

$$\lim_{T\to\infty}\sqrt{T}\operatorname{corr}(\hat{e}, e) = p\{(2+\gamma_4)/(2p+\Delta)\}^{1/2},$$

where $\gamma_4 = \kappa_4/\sigma^4$ *is the kurtosis of* ε_t *and* $\Delta = \kappa_4\sum_{t=0}^{\infty}\{\xi'(t)\Gamma^{-1}\xi(t)\}^2$.

Proof. First of all, since the ε_t's are iid,

$$\operatorname{var}\left(\sum_{t=p+1}^{T}\varepsilon_t^2\right) = (T-p)(2\sigma^4 + \kappa_4). \tag{6}$$

Using the spectral representations, we notice that the only difference between the Gaussian and non-Gaussian cases is the appearance of the term $K(f_4)$ in Theorem 1. Thus an argument similar to that leading to Theorem 3 shows that

$$\operatorname{cov}(\tilde{X}_T'\Gamma^{-1}\tilde{X}_T, \varepsilon_t^2) = 2\sigma^4\xi'(T-t)\Gamma^{-1}\xi(T-t) + K_t(f_4),$$

where

$$K_t(f_4) = \int_{R_1\cup R_2\cup R_3} e'(\omega)\Gamma^{-1}e(\lambda)e^{-i\omega t}e^{-i\lambda t}A(e^{-i\omega})A(e^{-i\lambda})f_4(\omega, \lambda, \mu)\,d\omega\,d\lambda\,d\mu.$$

By Theorem 2, it is easy to verify that

$$K_t(f_4) = \kappa_4\xi'(T-t)\Gamma^{-1}\xi(T-t).$$

Hence

$$\text{cov}\left(\tilde{X}_T'\Gamma^{-1}\tilde{X}_T,\ \sum_{t=p+1}^{T}\varepsilon_t^2\right) = p(2\sigma^4+\kappa_4)/\sigma^2 + o(1). \tag{7}$$

Likewise, we can show that

$$\text{var}(\tilde{X}_T'\Gamma^{-1}\tilde{X}_T) = 2p+\Delta, \tag{8}$$

where

$$\Delta = \int_{R_1\cup R_2\cup R_3} \{e'(\omega)\Gamma^{-1}e(\lambda)\}\cdot\{e'(\mu)\Gamma^{-1}e(-\omega-\lambda-\mu)\} \cdot A(e^{-i\omega})A(e^{-i\lambda})A(e^{-i\mu})A(e^{i(\omega+\lambda+\mu)})f_4(\omega,\lambda,\mu)\,d\omega\,d\lambda\,d\mu.$$

Using Theorem 2, after some tedious calculation, we can show that

$$\Delta = \kappa_4\sum_{t=0}^{\infty}\{\xi'(t)\Gamma^{-1}\xi(t)\}^2.$$

The conclusion follows from (6), (7), and (8). □

5. Discussion and concluding remarks

The results established in previous sections suggest that assessing the conditional prediction error is a difficult problem. Throughout this article, we have been using e, the conditional prediction error, as the measure of prediction power. One can also use the unconditional prediction error, namely $E(e)$, for the same purpose. It is our point of view, however, that the conditional version is more relevant in practice because it is a measure of how well one can do given the data at hand. The unconditional version makes sense only when the data are replicable.

One obvious observation is that as $T\to\infty$, $\text{corr}(\hat{e},e)$ tends to zero. In other words, $\hat{e}$ and e are asymptotically uncorrelated! Moreover, one is not likely to find any natural estimate statistic $\hat{e}$ that has positive asymptotic correlation with e. This suggests that assessing mean square prediction error in autoregressive models is nearly impossible. Taking one step back, we can evaluate the second-order term of $\text{corr}(\hat{e},e)$. From Theorem 4, the second-order term of $\text{corr}(\hat{e},e)$ is

$$C = p\{(2+\gamma_4)/(2p+\Delta)\}^{1/2}.$$

Hence the larger the C is, the larger $\text{corr}(\hat{e},e)$ is. We close this article by making the following remarks.

Remark 1. Other things being equal, C is increasing with respect to γ_4, the kurtosis of the error distribution. In other words, $\text{corr}(\hat{e},e)$ is large when the error distribution has heavy tails.

Remark 2. Other things being equal, C is decreasing with respect to Δ. The smaller the Δ is, the larger $\text{corr}(\hat{e},e)$ is. It is easy to show that $\Delta\to 0$ is equivalent to $\lambda_{\min}(\Gamma)\to\infty$ where $\lambda_{\min}(\Gamma)$ is the smallest eigenvalue of the matrix Γ. For the $AR(1)$ model, $\Gamma=\sigma^2/(1-\alpha^2)$ is a scalar and $\lambda_{\min}(\Gamma)\to\infty$ is equivalent to $|\alpha|\to 1$. For the general $AR(p)$ model, it can be shown [3,

p. 88] that $\lambda_{\min}(\Gamma) \to \infty$ if and only if $|\phi_i| \to 1$, $i = 1, \ldots, p$. In other words, all the roots have to be near the unit circle, a much stronger condition than near nonstationarity.

Remark 3. The previous two remarks imply that $\operatorname{corr}(\hat{e}, e)$ is large only for extremely non-Gaussian or nearly nonstationary processes. This, however, does not mean that prediction errors for such processes are easy to estimate. On the contrary, for such processes both e and $\hat{e}$ are very large and accurate prediction is out of question in the first place. Another interesting situation occurs when p is large. For Gaussian processes, Theorem 3 shows that $\operatorname{corr}(\hat{e}, e)$ is increasing as p increases. This seems to imply that higher-dimensional models become preferable, a violation of the widely held parsimony principle.

We are facing a serious dilemma here. On the one hand, it is desirable to estimate e rather than $E(e)$ for practical purposes. On the other hand, estimating e seems nearly impossible. The use of the correlation coefficient is quite natural because when an estimated prediction error is small one is inclined to think that e is also small, or that the model can produce accurate prediction. Various model selection criteria are derived based on this idea. The results that we have obtained indicate that this line of thinking has serious problems.

ACKNOWLEDGMENT

The authors are grateful to a referee for pointing out an error in §4.

REFERENCES

1. H. Akaike, *Statistical predictor identification*, Ann. Inst. Statist. Math. **22** (1970), 203–217.
2. ———, *Information theory and an extension of the maximum likelihood principle*, Proc. 2nd International Symposium on Information Theory (B. N. Petrov and Csáki, eds.), Akademiai Kiado, Budapest, 1973, pp. 267–281.
3. T. W. Anderson and R. P. Mentz, *On the structure of the likelihood function of autoregressive and moving average models*, J. Time Ser. Anal. **1** (1980), 83–94.
4. D. R. Brillinger, *Some history of the study of higher-order moments and spectra*, Statistica Sinica **1** (1991), 465–476.
5. W. A. Fuller and D. P. Hasza, *Properties of predictors for autoregressive time series*, J. Amer. Statist. Assoc. **76** (1981), 155–161.
6. I. Johnstone, *On inadmissibility of some unbiased estimates of loss*, Statistical Decision Theory and Related Topics IV, vol. 1 (S. S. Gupta and J. O. Berger, eds.), Springer-Verlag, New York, 1988, pp. 361–379.
7. H. Linhart and W. Zucchini, *Model selection*, Wiley, New York, 1986.
8. P. Shaman, *Properties of estimates of the mean square error of prediction in autoregressive models*, Studies in Econometrics, Time Series, and Multivariate Statistics (S. Karlin, T. Amemiya, and L. A. Goodman, eds.), Academic Press, New York, 1983, pp. 331–342.
9. P. Shaman and R. A. Stine, *The bias of autoregressive coefficient estimators*, J. Amer. Statist. Assoc. **83** (1988), 842–848.
10. R. Shibata, *Selection of the order of an autoregressive model by Akaike's information criterion*, Biometrika **63** (1976), 117-126.
11. T. P. Speed, *What is an analysis of variance (with discussion)?*, Ann. Statist. **15** (1987), 885–941.

DEPARTMENT OF STATISTICS, THE WHARTON SCHOOL OF THE UNIVERSITY OF PENNSYLVANIA, PHILADELPHIA, PENNSYLVANIA 19104-6302