An algorithm for robust fitting of autoregressive models

Dimitris N. Politis*

Abstract: An algorithm for robust fitting of AR models is given, based on a linear regression idea. The new method appears to outperform the Yule-Walker estimator in a setting of data contaminated with outliers.

Keywords: ARMA models, linear time series, outliers. **JEL Classification:** C1; C5.

1 Introduction

Consider data $X_1, \ldots, X_n$ arising as a stretch of a second-order stationary time series $\{X_t, t \in \mathbb{Z}\}$ with autocovariance sequence $\gamma_k = Cov(X_t, X_{t+k})$. We will assume that $EX_t = 0$ which, from a practical point of view, means that the data have been de-trended. An autoregressive model of order p, i.e., an AR(p), is defined by the following recursion:

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t, \text{ for all } t \in \mathbb{Z}, \tag{1}$$

where Z_t is a second-order stationary white noise with $EZ_t = 0$ and $EZ_t^2 = \sigma^2$. We will assume the above AR(p) model is *causal*, i.e., that for any $k > 0$, Z_{t+k} is uncorrelated to $\{X_{t-s}, s \geq 0\}$; see Brockwell and Davis (1991). Multiplying both sides of (1) by X_{t-k} and taking expectations, we derive:

$$\gamma_k = \phi_1 \gamma_{k-1} + \ldots + \phi_p \gamma_{k-p} + \sigma^2 \delta_k, \text{ for all } k \geq 0 \tag{2}$$

where $\delta_k = 0$ for $k \neq 0$ but $\delta_0 = 1$.

Plugging the sample autocovariance $\hat{\gamma}_k = n^{-1} \sum_{i=1}^{n-|k|} X_t X_{t+|k|}$ in place of the true γ_k in eq. (2) for $k = 0, 1, \ldots, p$, the well-known Yule-Walker (YW) equations are derived:

$$\hat{\gamma}_k = \phi_1 \hat{\gamma}_{k-1} + \ldots + \phi_p \hat{\gamma}_{k-p} + \sigma^2 \delta_k, \text{ for } \ k = 0, \ldots, p. \tag{3}$$

whose unique solution $\hat{\phi}_1, \ldots, \hat{\phi}_p$ and $\hat{\sigma}^2$ forms the well-known YW estimator that is is asymptotically efficient in the context of a Gaussian AR series. Nevertheless, the YW estimator loses its asymptotic efficiency under a non-Gaussian distributional assumption; see e.g. Sengupta and Kay (1989).

In what follows, we describe a simple estimation algorithm for AR model fitting; it is not a fast algorithm but it is promising in improving the finite-sample accuracy of the YW estimators when outliers are present. The new algorithm exemplifies robustness against outliers, and in particular against clusters of (two or more) outliers.

*Department of Mathematics, and Economics, Univ. of California at San Diego, La Jolla, CA 92093-0112, USA; tel.: (858) 534-5861, fax: (858) 534-5273, e-mail: dpolitis@ucsd.edu. Research partially supported by NSF grant SES-04-18136.

2 Motivation of the new algorithm

In what follows, the focus is on fitting an AR(p) model; for concreteness, the order p is assumed known. Let $\underline{\phi}_p = (\phi_1, \ldots, \phi_p)'$, $\underline{\gamma}_k = (\gamma_1, \ldots, \gamma_k)'$, and $\underline{\hat{\gamma}}_k = (\hat{\gamma}_1, \ldots, \hat{\gamma}_k)'$. Recall that the YW estimator of $\underline{\phi}_p$ and σ^2 is a (linear) function of $\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_p$; it is asymptotically efficient if the series is Gaussian AR(p) in which case $(\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_p)$ is (approximately) a sufficient statistic for $\underline{\phi}_p$ and σ^2. Therefore, in the Gaussian case, one is justified to just look at functions of $(\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_p)$.

Nevertheless, $(\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_p)$ is not necessarily sufficient without Gaussianity. It seems natural that in the context of second-order stationarity, a general estimator of $\underline{\phi}_p$ and σ^2 would be a function of all the second-order information available, i.e., $\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_{n-1}$. Since $\hat{\gamma}_k$ is unreliable for large k, i.e., when $n - k$ is small, it makes sense to base our estimator on $\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_{p'}$ where p' is potentially large as compared to p, but small as compared to n; in particular, we require that $n - p'$ is large, i.e., that

$$p \leq p' \leq cn \quad \text{for some} \quad c \in (0, 1). \tag{4}$$

Asymptotically, any value of $c \in (0, 1)$ guarantees that $n - cn \to \infty$ when $n \to \infty$ but for practical sample sizes of the order of 100 or 1,000 a reasonable choice for c must be small, say in the interval $[0.1, 0.2]$, for $n - cn$ to be large. An easy way to satisfy (4) is to let $p' = \max(p, \lfloor cn \rfloor)$ with c as above.

As eq. (2) implies the YW eq. (3) by plugging in $\hat{\gamma}_k$ for γ_k, the same eq. (2) also implies the 'Extended' Yule-Walker (EYW) equations:

$$\hat{\gamma}_k = \phi_1 \hat{\gamma}_{k-1} + \ldots + \phi_p \hat{\gamma}_{k-p} + \sigma^2 \delta_k, \text{ for } \quad k = 0, \ldots, p'. \tag{5}$$

3 The notion of robustness

It is desirable to have estimators that are robust to 'outliers', i.e., data points whose value is extreme compared to the bulk of the data. Outliers are generally due either to contaminated/corrupted data, or to heavy-tailed error distributions. For discussion on robustness see Franke et al. (1984), and Hampel et al. (1986).

To give an example, consider the Gaussian AR(1) model $X_t = \phi_1 X_{t-1} + Z_t$ with Z_t i.i.d. N(0,1). Suppose $n = 200$, and that the 100th observation has been corrupted by an outlier resulting into $X_{100} = B$. All $\hat{\gamma}_k$ are subsequently corrupted; instead of the usual $\hat{\gamma}_k = \gamma_k + O_P(1/\sqrt{n})$ we now have $\hat{\gamma}_k = \gamma_k + O_P(1/\sqrt{n}) + O(B/n)$. The situation is further aggrevated if a second outlier is found close to the first one, say at the 101th observation. So assume X_{101} is of the order of magnitude of B (positive or negative); consequently,

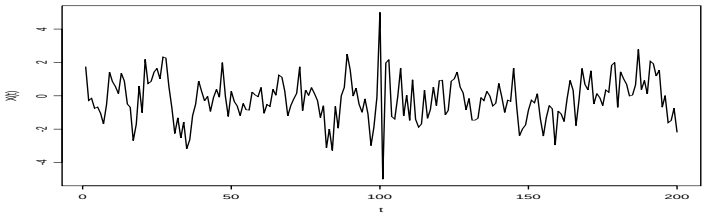

Figure 1: Plot of Gaussian AR(1) series $X_1, \ldots, X_{200}$.

estimation of the lag-0 and lag-1 autocovariances is adversely affected more than the others since $\hat{\gamma}_k = \gamma_k + O_P(1/\sqrt{n}) + O(B/n)$ for $k > 1$, but $\hat{\gamma}_k = \gamma_k + O_P(1/\sqrt{n}) + O(B^2/n)$ for $k = 0$ or 1. Hence, the YW estimator $\hat{\phi}_1 = \hat{\gamma}_1/\hat{\gamma}_0 = \phi_1 + O_P(1/\sqrt{n}) + O(B^2/n)$.

Figure 1 shows a plot of a Gaussian AR(1) series with $\phi_1 = 1/2$, and $n = 200$. The corrupted values are $X_{100} = 5$ and $X_{101} = -5$. The estimated autocovariances up to lag 10 were: $(\hat{\gamma}_0, \hat{\gamma}_1 \ldots, \hat{\gamma}_{10}) = (1.728, 0.517, 0.379, 0.263, 0.155, -0.075, -0.241, -0.121, -0.157, -0.195, -0.131)$, and the YW estimator is $\hat{\phi}_1 = \hat{\gamma}_1/\hat{\gamma}_0 = 0.517/1.728 = 0.30$. If it were not for the presence of back-to-back outliers, this estimator would be bigger by a term of order $B^2/n = 0.125$. In particular, for the uncorrupted dataset, the estimate of ϕ_1 would be about $(0.517 + 0.125)/(1.728 - 2 \cdot 0.125) \simeq 0.43$.

Note that the EYW eq. (5) for $k = 2$ gives the alternative estimator for ϕ_1 as $\hat{\gamma}_2/\hat{\gamma}_1 = 0.73$. Additional valid estimators for ϕ_1 are given by $\hat{\gamma}_3/\hat{\gamma}_2$, $\hat{\gamma}_4/\hat{\gamma}_3$, $\hat{\gamma}_5/\hat{\gamma}_4$, etc. It should be possible to combine all the above valid estimates of ϕ_1 to get a more accurate estimator. A way to do this is suggested by the observation that eq. (5) for $k = 1, 2, \ldots, p'$ represents points on the scatterplot of $\hat{\gamma}_k$ vs. $\hat{\gamma}_{k-1}$. For example, all points in the scatterplot of Figure 2. would—if all $\hat{\gamma}_k$ were close to the true value γ_k—be close to the straight line passing through the origin with slope ϕ_1.

This motivates the following idea: run a straight line regression on this scatterplot (with no intercept term) to estimate this slope. Figure 3 shows a plot of this regression-type estimate of ϕ_1 for different values of p'. It is apparent that using any p' bigger than p is beneficial; in particular, a high value of p' is desirable provided $n - p'$ is large as well as eq. (4) ensures. Also apparent is a certain insensitivity of the estimator on the value of p' as long as the latter is in the right range.

Figure 2 sheds intuition on why the choice of p' is not a crucial matter. The points on the scatterplot for high values of k are crowded near the origin by virtue of the exponential decay of γ_k. But eq. (5) describes a linear regression with no intercept term; thus, the fitted line will necessarily go through the origin, and points that are near the origin will

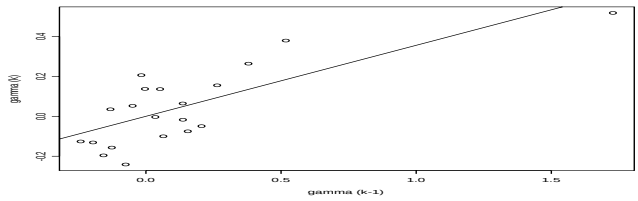

Figure 2: Scatterplot of $\hat{\gamma}_k$ vs. $\hat{\gamma}_{k-1}$ for $k = 1, \ldots, p'$ with $p' = 20$ and LS line superimposed.

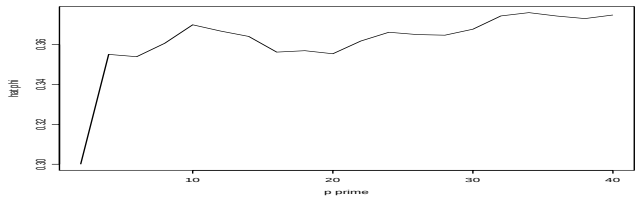

Figure 3: LS estimate of ϕ_1 as function of p'.

consequently have negligible 'leverage'.

4 The algorithm

Let the $p' \times p$ matrix $\mathbf{\Gamma}_{p',p}$ with jth column given by the vector $(\gamma_{1-j}, \gamma_{2-j}, \ldots, \gamma_{p'-j})'$. By eq. (2), the matrix equation follows:

$$\underline{\gamma}_{p'} = \mathbf{\Gamma}_{p',p}\underline{\phi}_p \tag{6}$$

Letting the $p' \times p$ matrix $\hat{\mathbf{\Gamma}}_{p',p}$ have jth column given by the vector $(\hat{\gamma}_{1-j}, \hat{\gamma}_{2-j}, \ldots, \hat{\gamma}_{p'-j})'$, we can write

$$\underline{\hat{\gamma}}_{p'} = \hat{\mathbf{\Gamma}}_{p',p}\underline{\phi}_p + \underline{\epsilon}. \tag{7}$$

that serves as definition for the the 'error' vector $\underline{\epsilon}$. Under weak assumptions, Brockwell and Davis (1991) show $\hat{\gamma}_k = \gamma_k + O_P(1/\sqrt{n})$ and $\hat{\rho}_k = \rho_k + O_P(1/\sqrt{n})$ for any $k \leq p'$ which is sufficient to ensure $\underline{\epsilon} = O_P(1/\sqrt{n})$. Here, $\rho_k = \gamma_k/\gamma_0$ and $\hat{\rho}_k = \hat{\gamma}_k/\hat{\gamma}_0$ are the true and sample lag-k autocorrelation respectively.

Eq. (7) can be viewed as a linear regression with 'response' $\underline{\hat{\gamma}}_{p'}$ and 'regressors' given by the columns of the matrix $\hat{\mathbf{\Gamma}}_{p',p}$. This is an example of regression with 'errors-in-variables'; see Fuller (1987), or Seber and Lee (2003). The usual LS solution[†] applies in this context

[†]A different regression technique such as L_1 or LAD (Least Absolute Deviations) or another robust

as well; thus, our proposed estimator of $\underline{\phi}_p$ is

$$\hat{\underline{\phi}}_{p',p} = \left(\hat{\mathbf{\Gamma}}'_{p',p}\hat{\mathbf{\Gamma}}_{p',p}\right)^{-1}\hat{\mathbf{\Gamma}}'_{p',p}\hat{\underline{\gamma}}_{p'}. \tag{8}$$

The special case $p' = p$ yields the YW estimator.

Lemma 1 *Fix a value $p' \geq p$.*
(a) If $\hat{\rho}_k \to \rho_k$ in probability for all $k = 1, \ldots, p'$, then $\hat{\underline{\phi}}_{p',p} \to \underline{\phi}_p$ in probability as $n \to \infty$.
(a) If $\hat{\rho}_k = \rho_k + O_P(1/\sqrt{n})$ for all $k = 1, \ldots, p'$, then $\hat{\underline{\phi}}_{p',p} = \underline{\phi}_p + O_P(1/\sqrt{n})$ as well.

Proof of Lemma 1. Although eq. (8) expresses $\hat{\underline{\phi}}_{p',p}$ as a function of the $\hat{\gamma}_k$'s, after dividing through by $\hat{\gamma}_0$ and performing the necessary cancellations, it is seen that $\hat{\underline{\phi}}_{p',p}$ is only a function of the sample autocorrelations $\hat{\rho}_k$ for $k = 1, \ldots, p'$. Furthermore, up to a set whose probability tends to zero, this function is continuously differentiable. Thus, part (a) follows from the continuous mapping theorem, while part (b) from the delta-method.□

Estimating σ^2 can be done in the usual way, i.e., plugging our estimator $\hat{\underline{\phi}}_{p',p}$ in the YW eq. (3) corresponding to $k = 0$ and solving for σ^2 to obtain

$$\hat{\sigma}^2_{p'} = \hat{\gamma}_0 - \hat{\underline{\gamma}}'_p\ \hat{\underline{\phi}}_{p',p}. \tag{9}$$

Lemma 2 *(a) If $\hat{\gamma}_0 \to \gamma_0$ and $\hat{\underline{\phi}}_{p',p} \to \underline{\phi}_p$ in probability, then $\hat{\sigma}^2_{p'} \to \sigma^2$ in probability as $n \to \infty$.*
(a) If $\hat{\gamma}_0 = \gamma_0 + O_P(1/\sqrt{n})$ and $\hat{\underline{\phi}}_{p',p} = \underline{\phi}_p + O_P(1/\sqrt{n})$, then $\hat{\sigma}^2_{p'} = \sigma^2 + O_P(1/\sqrt{n})$ as well.

In conclusion, note that the EYW eq. (5) were also used by Cadzow (1982) in formulating some alternative estimation procedures; in the case of AR models, Cadzow (1982) proposed a constrained Maximum Likelihood approach, as well as an approximate Singular Value Decomposition method. By comparison, our estimator $\hat{\underline{\phi}}_{p',p}$ enjoys extreme computational simplicity using the linear system (8). Furthermore, the performance of estimator $\hat{\underline{\phi}}_{p',p}$ does not seem to hinge on an 'optimal' choice of the value p', i.e., $\hat{\underline{\phi}}_{p',p}$ will perform well for a whole range of possible p' values as suggested in Section 2.

regression might be employed here in place of L_2 regression (LS); see e.g. Hampel et al. (1986), or Seber and Lee (2003). The disadvantage of robust regression is lack of a closed form solution such as (8); we will not pursue this further here for lack of space.

5 A finite-sample simulation

We revisit the contaminated AR(1) example of Section 3; six contamination patterns were considered:
(I) $B = 5, b = 5$; (II) $B = 5, b = 0$; (III) $B = 0, b = 5$; (IV) $B = 0, b = 0$; (V) $B = -5, b = 5$; (VI) $B = -5, b = 0$. The columns of Tables 1 and 2 correspond to: (i): $\hat{\phi}_{p,p}$, (ii): $\hat{\phi}_{n/10,p}$, (iii): $\hat{\phi}_{n/5,p}$, (iv): $\hat{\phi}_{n/2,p}$, (v): $\hat{\sigma}^2_{p'=p}$, (vi): $\hat{\sigma}^2_{p'=n/10}$, (vii): $\hat{\sigma}^2_{p'=n/5}$, and (viii): $\hat{\sigma}^2_{p'=n/2}$.

The findings in Tables 1 and 2 are encouraging. First, the robust estimator $\hat{\sigma}^2_{p'}$ appears to be uniformly better than the YW estimator $\hat{\sigma}^2_{p'=p}$ in all cases except in the uncontaminated case (IV) where the YW estimator is asymptotically efficient. Furthermore, the robust estimator $\hat{\phi}_{p',p}$ dominates the YW estimator $\hat{\phi}_{p,p}$ in all cases of Table 1 except (I) and (IV); the same is true for Table 2 when p' is chosen in the range $[n/10, n/5]$.

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
(I)	.105	.132	.125	.117	.787	.760	.743	.702
(II)	.237	.234	.224	.212	.621	.550	.540	.519
(III)	.245	.237	.229	.213	.643	.568	.557	.536
(IV)	.152	.156	.152	.147	.183	.184	.185	.188
(V)	.509	.501	.493	.477	1.28	1.19	1.19	1.18
(VI)	.226	.222	.210	.196	.607	.538	.527	.511

Table 1. Empirical Root Mean Squared Error (RMSE) of estimation in a contaminated Gaussian AR(1) model with $\phi_1 = 0.5$, $\sigma^2 = 1$, and $n = 50$; the contamination amounts to $X_{25} = B$ and $X_{26} = b$, and the empirical RMSEs are based on 333 replications.

Case (I) is a bit of a fluke, and deserves special discussion. To do this, let $\tilde{\gamma}_k$ denote the lag-k sample autocovariance from the uncontaminated—and unobserved—dataset. Then $\hat{\gamma}_0 \simeq \tilde{\gamma}_0 + (B^2 + b^2)/n$, and $\hat{\gamma}_1 \simeq \tilde{\gamma}_1 + (Bb)/n$. Consequently,

$$\hat{\phi}_1 = \hat{\rho}_1 = \frac{\hat{\gamma}_1}{\hat{\gamma}_0} \simeq \frac{\tilde{\gamma}_1 + (Bb)/n}{\tilde{\gamma}_0 + (B^2 + b^2)/n}$$

and thus $\hat{\phi}_1 \to 1/2$ when $B = b$ and $b \to \infty$. The above limit is an indication that the effect of back-to-back positive outliers is to bias the YW estimator $\hat{\phi}_1$ towards the value $1/2$. If the latter happens be the true value—as in our example—, the YW estimator is bestowed with a super-efficiency property.

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
(I)	.060	.066	.070	.092	.213	.201	.186	.160
(II)	.084	.081	.079	.089	.176	.153	.142	.126
(III)	.083	.076	.073	.084	.189	.164	.153	.135
(IV)	.063	.065	.071	.095	.100	.103	.106	.115
(V)	.194	.156	.141	.126	.402	.361	.350	.333
(VI)	.085	.076	.076	.087	.175	.150	.139	.124

Table 2. Same as Table 1 with $n = 200$, and contamination pattern: $X_{100} = B$ and $X_{101} = b$.

Returning to the uncontaminated case (IV), note that in the small-sample setting of Table 1, $\hat{\phi}_{p',p}$ appears competitive to $\hat{\phi}_{p,p}$ which by itself is a remarkable fact. In the large-sample setting of Table 2, the asymptotic efficiency of the YW estimator starts to kick in, but still $\hat{\phi}_{p',p}$ with $p' = n/10$ seems reasonably competitive.

Finally, our empirical work confirms that $\underline{\hat{\phi}}_{p',p}$ perform wells for a whole range of possible p' values of type $p' = \max(p, \lfloor cn \rfloor)$ with c in $[0.1, 0.2]$ as suggested in Section 2.

References

[1] Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods, 2nd ed.*, Springer, New York.

[2] Cadzow, J.A. (1982). Spectral estimation: an overdetermined rational model equation approach, *Proc. IEEE*, vol. 70, no. 9, 907–939.

[3] Franke, J., Härdle, W., and Martin, D. (1984). *Robust and nonlinear time series analysis*, Lecture Notes in Statistics vol. 26, Springer, New York.

[4] Fuller, W.A. (1987). *Measurement Error Models*, Wiley, New York.

[5] Hampel, F.R., Ronchetti, E., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics : The Approach Based on Influence Functions,* Wiley, New York.

[6] Seber, G.A.F. and Lee, A.J. (2003). *Linear Regression Analysis, 2nd Ed.*, Wiley, New York.

[7] Sengupta, D. and Kay, S. (1989). Efficient estimation of parameters for non-Gaussian autoregressive processes, *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 37, no. 6, pp. 785–794.