Predicting ARMA Processes

Overview

Prediction of ARMA processes resembles in many ways prediction in regression models, at least in the case of AR models. We focus on linear predictors, those that express the prediction as a weighted sum of past observations.

1. ARMA models, notation
2. Best linear predictor
3. Levinson's algorithm
4. Prediction errors
5. Discussion

ARMA processes

Review notation A *stationary* solution $\{X_t\}$ (or if its mean is not zero, $\{X_t - \mu\}$) of the linear difference equation

$$\begin{aligned} X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} &= w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \quad w_t \sim WN(0, \sigma^2) \\ \phi(B) X_t &= \theta(B) w_t \end{aligned} \tag{1}$$

In general, I will treat $\mu = 0$ (until we fit models to data).

Moving average The one-sided MA representation is

$$X_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B) w_t \; , \tag{2}$$

Autoregression The corresponding AR representation (assuming invertibility) is

$$X_t = w_t + \sum_{j=1}^{\infty} \pi_j X_{t-j} \;\text{ or } \pi(B) X_t = w_t \; . \tag{3}$$

Best linear predictor

Conditional mean Consider finding the best estimator of Y given that we can use *any* function of the observed variables $X_{1:n} = X_1, X_2, \ldots, X_n$ (where "best" means minimal mean squared error loss),

$$\min_g \mathbb{E}\,(Y - g(X_1, X_2, \ldots, X_n))^2 \;.$$

The answer is given by setting g to the conditional expected value of Y, $g(X_{1:n}) = \mathbb{E}\,Y|X_{1:n}$.

The proof resembles those used in regression analysis because we can think of the conditional expected value as a projection onto $X_{1:n}$. Add and subtract $\mathbb{E}\,(Y|X_{1:n})$, expand the square, and then observe that $\mathbb{E}\,(Y - \mathbb{E}\,(Y|X_{1:n})(\mathbb{E}\,(Y|X_{1:n}) - g(X_{1:n})) = 0$ (use the law of total expectation, $\mathbb{E}\,Y = \mathbb{E}_x \mathbb{E}_{y|x} Y$).

Best linear predictor In general, the conditional mean is a nonlinear function of $X_{1:n}$, but we will emphasize finding linear predictors for two reasons.

- In the Gaussian case, the conditional mean is linear.
- Linear predictors only require second-order properties of the process. Since we assume second-order stationarity, we can estimate these by averaging over time in the observed data.

We define (see **Property 3.3**) the best linear predictor of X_{n+m} (*i.e.*, m periods beyond the end of the observed time series) as $\hat{X}_{n+m}$ (the book writes this as X^n_{n+m})

$$\min_\alpha \mathbb{E} \left(X_{t+m} - (\hat{X}_{n+m} = \sum_{j=1}^{n} \alpha_j X_{n+1-j}) \right)^2 \tag{4}$$

Equivalently, we can define the best squared-error predictor by demanding orthogonal prediction errors,

$$\mathbb{E}\,(X_{n+m} - \hat{X}_{n+m}) X_j = 0, \quad j = 1, 2, \ldots, n \tag{5}$$

Yule-Walker equations, again Consider predicting one-step ahead at X_{n+1}. Write the coefficients in the form $\phi_{mk} = \phi_{\text{size,index}}$ (some books do these in the other order, so read carefully). Multiplying and taking expectations in

$$\mathbb{E}\, X_{n+1-k}(X_{n+1} - \sum \phi_{nj} X_{n+1-j}) = 0 \quad k = 1, \ldots, n \ ,$$

gives the $n \times n$ system of equations

$$\boldsymbol{\gamma} = \Gamma_n \boldsymbol{\phi}, \quad [\Gamma_n]_{ij} = \gamma(i-j). \tag{6}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)'$. Solving directly gives a large inverse and the solution to the prediction problem:

$$\hat{X}_{n+1} = \sum_j \phi_j X_{n+1-j} \ \text{ where } \boldsymbol{\phi} = \Gamma_n^{-1} \boldsymbol{\gamma} \ .$$

In matrix form, the mean squared error of one-step ahead prediction reduces to

$$\begin{aligned} \mathbb{E}\,(X_{n+1} - \hat{X}_{n+1})^2 &= \mathbb{E}\,(X_{n+1} - \boldsymbol{\phi}' X_{1:n})^2 \\ &= \gamma(0) - 2\,\boldsymbol{\phi}'\boldsymbol{\gamma} + \boldsymbol{\phi}'\Gamma_n\boldsymbol{\phi} \\ &= \gamma(0) - \boldsymbol{\gamma}'\,\Gamma_n^{-1}\,\boldsymbol{\gamma}\ , \end{aligned} \tag{7}$$

when you substitute $\boldsymbol{\phi} = \Gamma_n^{-1}\boldsymbol{\gamma}$. The expression (7) resembles the expression in a linear regression for the residual sum of squares, $RSS = y'y - \hat{\beta}'(X'X)\hat{\beta}$.

Note: the problem changes slightly when the forecast horizon increases from predicting X_{n+1} to X_{n+m} for $m > 1$; the covariance vector $\boldsymbol{\gamma}$ changes from $(\gamma_1, \ldots, \gamma_n)'$ to $(\gamma_m, \ldots, \gamma_{n+m-1})'$.

Levinson's recursion

Problem How do we solve the prediction equations (6), which in general concern an $n \times n$ system of equations? It turns out that there is a very nice *recursive* solution.

Levinson's recursion (p 112 or 113) takes as input $\gamma(0), \gamma(1), \ldots$ and provides the coefficients $\phi_{k1}, \phi_{k2}, \ldots, \phi_{kk}$ of the AR(k) model that minimizes the MSE

$$\min \mathbb{E}\,(X_{n+1} - \phi_{k1}X_n - \phi_{k2}x_{n-1} - \cdots - \phi_{kk}X_{n-k+1})^2$$

and also gives the MSE itself (denoted P in the text)

$$\sigma_k^2 = \mathbb{E}\left(X_{n+1} - \phi_{k1}X_n - \phi_{k2}X_{n-1} - \cdots - \phi_{kk}X_{n-k+1}\right)^2 .$$

Along the way to producting the solution ϕ_{kj}, the recursion also solves the lower order approximations of order $p = 1, 2, \ldots, k-1$.

Algorithm Initialize $\phi_{00} = 0$ and $\sigma_0^2 = \gamma(0) = \mathrm{Var}(X_t)$. Compute the *reflection coefficient* ϕ_{kk} (which gives the PACF) using ($k = 1, 2, \ldots$)

$$\phi_{kk} = \frac{\rho(k) - \sum_{j=1}^{k-1}\phi_{k-1,j}\rho(k-j)}{1 - \sum_{j=1}^{k-1}\phi_{k-1,j}\rho(j)}$$

(Note that $\phi_{11} = \rho(1)$.) The update to the prediction MSE is

$$\sigma_k^2 = \sigma_{k-1}^2(1 - \phi_{kk}^2) .$$

Since the asolute value of the reflection coefficient $|\phi_{kk}| < 1$, it follows that the error variances are decreasing, $\sigma_k^2 \le \sigma_{k-1}^2$. (Or, since the minimization over more parameters cannot give a larger MSE, maybe this is a way to prove $|\phi_{kk}| \le 1$!)

The remaining coefficients that determine the predictor are

$$\phi_{kj} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j} .$$

Derivation resembles the updating a regression equation when a variable is added to the fitted model. The special form relies on the symmetry of Γ_k around both the usual and transverse diagonal. Write the k prediction equations that determine $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kk})$ in correlation form as

$$\begin{pmatrix} R_{k-1} & \tilde{\boldsymbol{\rho}}_{k-1} \\ \tilde{\boldsymbol{\rho}}'_{k-1} & 1 \end{pmatrix} \begin{pmatrix} \beta \\ \phi_{kk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\rho}_{k-1} \\ \rho(k) \end{pmatrix}$$

where R_k is the $k \times k$ correlation matrix of the process, $\boldsymbol{\rho}_k = (\rho(1), \ldots, \rho(k))$, $\tilde{\boldsymbol{\rho}}_k = (\rho(k), \rho(k-1), \ldots, \rho(1))'$ (the elements in ρ_k in reversed order) and the leading $k-1$ terms in $\boldsymbol{\phi}_k$ are $\beta' = (\phi_{k1}, \phi_{k2}, \ldots, \phi_{k,k-1})$. Write this system of equations as two equations, one a matrix equation and the other a scalar equation,

$$R_{k-1}\beta + \tilde{\boldsymbol{\rho}}_{k-1}\phi_{kk} \quad = \quad \boldsymbol{\rho}_{k-1}$$

$$\tilde{\boldsymbol{\rho}}'_{k-1}\beta \quad + \quad \phi_{kk} \quad = \quad \rho(k)$$

Noting that R_{k-1} is invertible (make sure you remember why), multiply the first equation by $\tilde{\boldsymbol{\rho}}'_{k-1}R^{-1}_{k-1}$ and combine the equations to eliminate β. Then solve for ϕ_{kk}, obtaining the equation for the reflection coefficient

$$\phi_{kk} = \frac{\rho(k) - \tilde{\boldsymbol{\rho}}'_{k-1}R^{-1}_{k-1}\boldsymbol{\rho}_{k-1}}{1 \;-\; \tilde{\boldsymbol{\rho}}'_{k-1}R^{-1}_{k-1}\tilde{\boldsymbol{\rho}}_{k-1}} = \frac{\rho(k) - \tilde{\boldsymbol{\rho}}'_{k-1}\boldsymbol{\phi}_{k-1}}{1 \;-\; \boldsymbol{\rho}_{k-1}{}'\boldsymbol{\phi}_{k-1}} \tag{8}$$

since $R^{-1}_{k-1}\boldsymbol{\rho}_{k-1} = \boldsymbol{\phi}_{k-1}$ (the coefficients obtained at the prior step) and $R^{-1}_{k-1}\tilde{\boldsymbol{\rho}}_{k-1} = \tilde{\boldsymbol{\phi}}_{k-1}$ (in reverse order; think about this one... you might want to consider the rotation matrix W for which $\tilde{\rho} = W\rho$. W has 1s along its opposite diagonal). Plugging this expression for ϕ_{kk} back in the first equation gives the formula for the leading $k-1$ terms:

$$\beta = \boldsymbol{\phi}_{k-1} - \phi_{kk}\tilde{\boldsymbol{\phi}}_{k-1}$$

Error variance To see this result, consider the usual regression model. In the linear model $y = x'\beta + \epsilon$ in which x is a random vector that is uncorrelated with ϵ (β is fixed), the variance of the error is given by

$$\mathrm{Var}(y) = \mathrm{Var}(x'\beta) + \mathrm{Var}(\epsilon) \Rightarrow \quad \sigma^2_\epsilon = \sigma^2_y - \beta'\,\mathrm{Var}(x)\beta$$

This expression suggests the relationship among the sums of squares, Total SS = Residual SS + Regression SS, or $Y'Y = e'e + \hat{\beta}'(X'X)\hat{\beta}$.

The unexplained variance after k steps of Levinson's recursion is thus

$$\begin{aligned} \sigma^2_k &= \gamma(0) - \boldsymbol{\phi}'_k\Gamma_k\boldsymbol{\phi}_k \\ &= \gamma(0)(1 - \boldsymbol{\rho}'_k\boldsymbol{\phi}_k) \\ &= \sigma^2_{k-1}(1 - \phi^2_{kk})\;. \end{aligned} \tag{9}$$

Since ϕ_{kk} is the partial correlation between X_t and X_{t-p} given X_{t-1},. . ., X_{t-p+1}, think about the last step as in the use of the R^2 statistic in regression analysis. Adding X_k to a model that contains the predictors $X_1, \ldots, X_{k-1}$ explains ϕ^2_{kk} (the square of its partial correlation with the response) of the remaining variation.

Algebraically, substitute partitioned vectors into the second expression for σ^2_k given in (9) and solve; it's not too messy:

$$1 - \boldsymbol{\rho}'_k\boldsymbol{\phi}_k \quad = \quad 1 - \boldsymbol{\rho}'_{k-1}\beta - \rho(k)\phi_{kk}$$

$$
\begin{aligned}
&= 1-\boldsymbol{\rho}'_{k-1}(\boldsymbol{\phi}_{k-1}-\phi_{kk}\tilde{\boldsymbol{\phi}}_{k-1})-\rho(k)\phi_{kk}\\
&= (1-\boldsymbol{\rho}'_{k-1}\boldsymbol{\phi}_{k-1})-\phi_{kk}(\rho(k)-\boldsymbol{\rho}'_{k-1}\tilde{\boldsymbol{\phi}}_{k-1})\\
&= (1-\boldsymbol{\rho}'_{k-1}\boldsymbol{\phi}_{k-1})-\phi_{kk}^2(1-\boldsymbol{\rho}'_{k-1}\boldsymbol{\phi}_{k-1})\\
&= (1-\boldsymbol{\rho}'_{k-1}\boldsymbol{\phi})k-1)(1-\phi_{kk}^2) \qquad (10)
\end{aligned}
$$

where the next-to-the-last line follows from (8).

Innovations algorithm alternatively solves recursively for the moving average representation, increasing the number of terms in the moving average form of the model. See page 115.

Prediction errors

Another view of predictor Levinson's algorithm (for AR models) and the corresponding innovations algorithm (MA models) determine the best linear predictor $\hat{X}_{n+m}$ and $\mathbb{E}\,(X_{n+m}-\hat{X}_{n+m})^2$ for a fixed lead m beyond the observed data. It is also useful to have expressions that summarize the effect of increasing m.

Prediction horizon The moving average representation (2) (from the innovations algorithm) is useful because the orthogonality of the errors. Write the time series $X_t=\sum\psi_j w_{t-j}$ in staggered form as

$$
\begin{array}{llllll}
X_{n+1} = & & & & w_{n+1} & +\psi_1 w_n+\psi_2 w_{n-1}+\cdots\\
X_{n+2} = & & & w_{n+2} & +\psi_1 w_{n+1} & +\psi_2 w_n+\psi_3 w_{n-1}+\cdots\\
X_{n+3} = & & w_{n+3} & +\psi_1 w_{n+2} & +\psi_2 w_{n+1} & +\psi_3 w_n+\psi_4 w_{n-1}+\cdots\\
X_{n+4} = & w_{n+4} & +\psi_1 w_{n+3} & +\psi_2 w_{n+2} & +\psi_3 w_{n+1} & +\psi_4 w_n+\psi_5 w_{n-1}+\cdots
\end{array}
$$

Since the white noise $w_t, w_{t-1},\ldots$ up to time n is "observable" given that we know the infinite past $X_{-\infty:n}$, the best linear predictor of X_{n+m} is

$$\hat{X}_{n+m}=\sum_{j=0}^{\infty}\psi_{m+j}w_{n-j}$$

Notice that the predictions are mean-reverting: $\hat{X}_{n+m}$ tends to $\mathbb{E}\,X_t=\mu$ as m increases.

Infinite past? This description of predictors and their MSE assumes we have the entire history of the process. This assumption is for convenience, and not unreasonable in practice. The convenience arises because in this setting the information in $X_n, X_{n-1}, \ldots$ is equivalent to that in $w_n, w_{n-1}, \ldots$ (the sigma fields agree). This equivalence allows us to swap between X_t and w_t.

For instance, consider predicting an ARMA(1,1) process (**Example 3.22**, p 119). The process is $X_t = \phi_1 X_{t-1} + w_t + \theta_1 w_{t-1}$. Clearly, the best predictor of X_{n+1} is

$$\hat{X}_{n+1} = \phi_1 X_n + \theta_1 w_n \ .$$

But if we observe only $X_{1:n}$, how can we learn w_n? We know from (3) that $w_n = \sum_j \pi_j X_{n-j}$, but this sum continues back in time past X_1. For a quick (and accurate so long as n is large relative to the strength of dependence) approximation to w_n, we can construct estimates of the errors from the start of the series. (Set $\tilde{w}_1 = 0$, then estimate $\tilde{w}_2 = X_2 - \phi_1 X_1 - \theta_1 \tilde{w}_1$ and continue recursively.)

MSE The mean squared prediction error is also evident in this expression,

$$\mathbb{E}\,(X_{n+m} - \hat{X}_{n+m})^2 = \sigma^2 \sum_{j=0}^{m-1} \psi_j^2 \ .$$

This prediction error approaches the variance of the process rapidly because typically only the leading ψ_j are large. For example, for an AR(1) process with $\phi_1 = 0.8$ and $\sigma^2 = 1$, $\text{Var}(X_t) = 1/(1-0.8^2) = 2.78$. The prediction MSE is

Lead	MSE
1	1
2	$1 + 0.8^2 = 1.64$
3	$1 + 0.8^2 + 0.64^2 = 2.05$
4	$1 + 0.8^2 + 0.64^2 + 0.512^2 = 2.31$

Good habit: plot the mean squared error $\sigma^2(\sum_{j=1}^k \psi_j^2)$ versus k to see how close the MSE has approached the series variance, $\text{Var}(X_t) = \sigma^2 \sum_j \psi_j^2$. For moderate values of k, the MSE is typically very near $\text{Var}(X_t)$, implying that the time series model predicts only slightly better than μ at this degree of extrapolation. ARMA models are most useful for short-term forecasts, particularly when you consider that this calculation gives an "optimistic" estimate of the actual MSE:

1. We don't know the infinite history;
2. We don't know the parameters ϕ, θ, μ;
3. We don't know the order of the process (p, q);
4. We don't even know that the process is ARMA.

When these are taken into account, it's likely that our MSE is larger than suggested by these calculations, perhaps higher than the MSE of simply predicting with $\overline{X}$.

Discussion

Example 3.23 illustrates the use of an ARMA model for forecasting the fish recruitment time series. Figure 3.6 shows the rapid growth of the MSE of an AR(2) forecast based on estimates. The forecasts are "interesting" for about six periods out and then settle down to the mean of the process.

Estimates? Up to now, we have considered the properties of ARMA processes. Now we have to see how well we can identify and then estimate these from data.