$\mathbf{ARMA}(p, q)$ Estimation and Selection

September 3, 2024

The material in this set of notes is based on S&S Chapter 3, specifically 3.6. We're going to start by learning how to **estimate** the $\mathbf{ARMA}(p, q)$ parameters, first as if we know p and q and then as if we do not.

Estimation

Method of Moments

These methods are based on our ability to relate the autocovariance function to the $\mathbf{ARMA}(p, q)$ parameters. Using whatever methods we have for recovering the unknown $\mathbf{ARMA}(p, q)$ parameters from the autocovariance function, we can obtain estimates of the $\mathbf{ARMA}(p, q)$ parameters by substituting the sample autocovariance function in for the true autocovariance function.

Yule-Walker Estimation of $\mathbf{AR}(p)$ Coefficients: The method of moments works best for the $\mathbf{AR}(p)$ model, in which case it is called **Yule-Walker** estimation of the $\mathbf{AR}(p)$ parameters coefficients. Remember, the $\mathbf{AR}(p)$ model is given by

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + w_t.$$

We know that the autocovariance function satisfies:

$$\gamma_y(0) = \phi_1 \gamma_y(1) + \cdots + \phi_p \gamma_y(p) + \sigma_w^2 \tag{1}$$

$$\gamma_y(h) = \phi_1 \gamma_y(h-1) + \cdots + \phi_p \gamma_y(h-p). \tag{2}$$

As long as we have $n \geq p$, we can plug in estimates of all of the autocovariance functions that appear in (1) to estimate σ_w^2. To estimate the remaining p coefficients $\phi_1, \dots, \phi_p$, we need p equations. We can restrict our attention to (1) and (2):

$$\begin{aligned} \gamma_y(1) &= \phi_1 \gamma_y(0) + \cdots + \phi_p \gamma_y(1-p) \\ &\vdots \\ \gamma_y(p) &= \phi_1 \gamma_y(p-1) + \cdots + \phi_p \gamma_y(0). \end{aligned}$$

This may look familiar - it's actually our forecasting equation for forecasting y_{p+1}! Rearranging gives:

$$\boldsymbol{A}_p \boldsymbol{\phi} = \boldsymbol{b}_p.$$

Replacing the true autocovariance functions with the sample autocovariance functions yields

$$\hat{\boldsymbol{A}}_p \hat{\boldsymbol{\phi}}_{YW} = \hat{\boldsymbol{b}}_p, \tag{3}$$

where $\hat{\boldsymbol{A}}_p$ has elements $\hat{a}_{p,ij} = \hat{\gamma}_y(i-j)$, $\hat{\boldsymbol{b}}_{p,i} = \hat{\gamma}_y(i)$, and $\hat{\boldsymbol{\phi}}_{YW}$ are the Yule-Walker estimators. Note that because computing $\hat{\boldsymbol{\phi}}_{YW}$ is the same as solving a forecasting equation, we can either solve (3) directly by inverting $\boldsymbol{A}_n$ or using the Durbin-Levinson algorithm.

If $\hat{\boldsymbol{\phi}}_{YW}$ and $\hat{\sigma}^2_{w,YW}$ are estimated from a time series $\boldsymbol{y}$ that is distributed according to a stationary $\mathbf{AR}(p)$ model, then as $n \to \infty$ (equivalently, as the time series gets longer):

- $\sqrt{n}\left(\hat{\boldsymbol{\phi}}_{YW} - \boldsymbol{\phi}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma_w^2 \boldsymbol{A}_p^{-1}\right)$;
- $\hat{\sigma}^2_{w,YW} \xrightarrow{p} \sigma_w^2$.

Furthermore, $\hat{\boldsymbol{\phi}}_{YW}$ will always be stationary.

Yule-Walker Estimation of MA(q) Coefficients: **Yule-Walker** estimation of the **MA**(q) parameter coefficients is much more challenging. Consider the **MA**(1) case:

$$y_t = \theta_1 w_{t-1} + w_t.$$

We know that the autocovariance function satisfies:

$$\gamma_y(0) = \left(1 + \theta_1^2\right) \sigma_w^2 \tag{4}$$

$$\gamma_y(h) = \begin{cases} \theta_1 \sigma_w^2 & h = 1 \\ 0 & h > 1 \end{cases} \tag{5}$$

Only $\gamma_y(0)$ and $\gamma_y(1)$ depend on the unknown **MA**(1) parameters. Plugging in the sample autocovariance values $\hat{\gamma}_y(0)$ and $\hat{\gamma}_y(1)$ yields the Yule-Walker equations for the **MA**(1) model. However, these Yule-Walker equations do not depend on the **MA**(1) parameters linearly as in the **AR**(p) case.

We can overcome the nonlinearity in the **MA**(1) case. We can obtain $\sigma_w^2 = \hat{\gamma}_y(1)/\theta_1$ from (5) and plug this into (4). Rearranging and dividing by the sample variance $\hat{\gamma}_y(0)$ yields

$$0 = \hat{\rho}_y(1)\theta_1^2 - \theta_1 + \hat{\rho}_y(1). \tag{6}$$

Then we can then solve (6) for θ_1 using the quadratic formula:

$$\theta_1 = \frac{1 \pm \sqrt{1 - 4\hat{\rho}_y(1)^2}}{2\hat{\rho}_y(1)}.$$

We can see that this will yield two real solutions when $|\hat{\rho}_y(1)| < 1/2$, in which case we choose the one that corresponds to an **invertible MA**(1) model.

When $|\hat{\rho}_y(1)| > 1/2$, there are no real solutions. This highlights an issue with method-of-moments estimation of **MA**(q) parameters that does not arise in method-of-moments estimation of **AR**(p) parameters. Certain autocovariance values may not be achievable under **MA**(q) models for any values of $\theta_1, \dots, \theta_q$ and σ_w^2, whereas all autocovariance values are achievable under an $\mathbf{AR}(p)$ model for some $\phi_1, \ldots, \phi_p$ and σ_w^2. This means that that it can be impossible to estimate $\mathbf{MA}(q)$ parameters for certain time series, if their sample autocovariance functions include values that cannot be achieved by an $\mathbf{MA}(q)$ model.

Challenges posed by nonlinearity are more evident if we add another moving average term and consider an $\mathbf{MA}\,(2)$ model,

$$y_t = \theta_1 w_{t-1} + \theta_2 w_{t-2} + w_t.$$

We know that the autocovariance function satisfies:

$$\gamma_y\,(0) = \left(1 + \theta_1^2 + \theta_2^2\right)\sigma_w^2 \tag{7}$$

$$\gamma_y\,(h) = \begin{cases} \theta_1\,(1+\theta_2)\,\sigma_w^2 & h = 1 \\ \theta_2\sigma_w^2 & h = 2 \\ 0 & h > 2 \end{cases} \tag{8}$$

Again, we can plug in the sample autocovariance values $\hat{\gamma}_y\,(0)$ and $\hat{\gamma}_y\,(1)$ to get the Yule-Walker equations for the $\mathbf{MA}(2)$ model:

$$\begin{aligned} \hat{\gamma}_y\,(0) &= \left(1 + \theta_1^2 + \theta_2^2\right)\sigma_w^2 \\ \hat{\gamma}_y\,(1) &= \theta_1\,(1+\theta_2)\,\sigma_w^2 \\ \hat{\gamma}_y\,(2) &= \theta_2\sigma_w^2. \end{aligned}$$

It is easy to see that these are very nonlinear in θ_1, θ_2, and σ_w^2 and difficult to solve.

Based on what we've seen so far, we can expect that solving the Yule-Walker equations will get more and more difficult as the order of the $\mathbf{MA}\,(q)$ model q increases. Furthermore, these problems will persist if we add autoregressive terms and Yule-Walker estimation of $\mathbf{ARMA}\,(p,q)$ parameters will be similarly intractable. We're going to need to take a different approach that does not rely on exact moment-matching instead, if we want to get method-of-moments estimates of $\mathbf{MA}\,(q)$ and $\mathbf{ARMA}(p,q)$ parameters. On top of the numerical challenges, estimates of $\mathbf{MA}\,(q)$ and $\mathbf{ARMA}(p,q)$ parameters obtained by solving the Yule-

Walker equations will be inefficient relative to the maximum likelihood estimates, i.e. they will tend to be more variable than the corresponding maximum likelihood estimators of the same quantities.

Moment-Based Estimation of ARMA(p, q) Coefficients via Innovations: Recall that another way of obtaining forecasts was via the innovations algorithm, which finds the coefficients $\boldsymbol{d}_m$ that minimize

$$v_m = \mathbb{E}\left[\left(y_{m+1} - \sum_{j=1}^{m} d_{mj}\left(y_{m+1-j} - \hat{y}_{m+1-j}\right)\right)^2\right].$$

The coefficients $d_{mj} \rightarrow \psi_j$ as $m \rightarrow \infty$. This is helpful, because it suggests that the coefficients when $m = n$ denoted by $\boldsymbol{d}_n$ will approximate ψ_j well, we can relate elements of ψ_j to $\phi_1, \dots, \phi_p$ and $\theta_1, \dots, \theta_q$. Remember that when our **ARMA**(p, q) model is stationary and invertible, we can write:

$$\phi(B)\psi(B)w_t = \theta(B)w_t.$$

Expanding this gives us our equations relating $\psi_1, \dots, \psi_n$ to $\phi_1, \dots, \phi_p$ and $\theta_1, \dots, \theta_q$.

$$\psi_j = \theta_j + \sum_{i=1}^{\min\{j,p\}} \phi_i \psi_{j-i} \text{ for } j = 1, \dots, q \tag{9}$$

$$\psi_j = \sum_{i=1}^{\min\{j,p\}} \phi_i \psi_{j-i} \text{ for } j = q+1, \dots, n \tag{10}$$

Because (10) does not involve θ_j at all and is linear in $\phi_1, \dots, \phi_p$, we can solve for the innovations estimates $\hat{\boldsymbol{\phi}}_I$ having plugged in $d_{n1}, \dots, d_{nn}$ for $\psi_1, \dots, \psi_n$. Then plugging $\hat{\boldsymbol{\phi}}_I$ for $\boldsymbol{\phi}$ and $d_{n1}, \dots, d_{nn}$ in for $\psi_1, \dots, \psi_n$, we can obtain $\hat{\boldsymbol{\theta}}_I$.

These estimators are not consistent, and inefficient relative to maximum likelihood estimators when estimating the parameters of **ARMA**(p, q) models with $q > 0$. Their advantage is that they are quite easy to compute, and that they can provide good starting values for more complicated algorithms for estimating **ARMA**(p, q) parameters, like those used to compute maximum likelihood estimators.

Maximum Likelihood Estimation

Maximum likelihood estimation of $\mathbf{ARMA}(p,q)$ also incorporates the distributional assumptions we made. Specifically, we assumed that the noise is independent and identically normally distributed, $w_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$. Because a vector $\boldsymbol{y}$ distributed according to an $\mathbf{ARMA}(p,q)$ model is linear in the noise, we know that $\boldsymbol{y}$ is normally distributed as well, with mean $\mathbb{E}[\boldsymbol{y}] = \mu \mathbf{1}_n$ and $\mathbb{V}[\boldsymbol{y}] = \boldsymbol{A}_n$, where $a_{n,ij} = \gamma_y(i-j)$. Letting $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ be $p \times 1$ and $q \times 1$ vectors of the autoregressive and moving average parameters in the $\mathbf{ARMA}(p,q)$ model, then we can write the likelihood of $\boldsymbol{y}$ as

$$p\left(\boldsymbol{y}|\boldsymbol{\phi}, \boldsymbol{\theta}, \mu, \sigma_w^2\right) = \frac{1}{\sqrt{2\pi\left|\boldsymbol{A}_n\right|}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y} - \mu\mathbf{1}_n\right)' \boldsymbol{A}_n^{-1}\left(\boldsymbol{y} - \mu\mathbf{1}_n\right)\right\}. \tag{11}$$

This is a quick, parsimonious way of writing the likelihood of $\boldsymbol{y}$, but as we saw with the forecasting algorithm, it can be computationally challenging to evaluate if n is large because each time we evaluate (11) for new values of the $\mathbf{ARMA}(p,q)$ parameters we need to invert and compute the determinant $\boldsymbol{A}_n$. Instead, we will derive the likelihood of $\boldsymbol{y}$ from the conditional distributions, making use of the fact that

$$\begin{aligned}
&p\left(\boldsymbol{y}|\boldsymbol{\phi}, \boldsymbol{\theta}, \mu, \sigma_w^2\right) = \\
&p\left(y_1|\boldsymbol{\phi}, \boldsymbol{\theta}, \mu, \sigma_w^2\right) p\left(y_2|y_1, \boldsymbol{\phi}, \boldsymbol{\theta}, \mu, \sigma_w^2\right) \dots p\left(y_n|y_{n-1}, \dots, y_1, \boldsymbol{\phi}, \boldsymbol{\theta}, \mu, \sigma_w^2\right).
\end{aligned}$$

Again, each conditional distribution $p\left(y_j|y_{j-1}, \dots, y_1, \boldsymbol{\phi}, \boldsymbol{\theta}, \mu, \sigma_w^2\right)$ will be normal, because each y_j is a linear function of normal noise variables. For this to be useful, though, we need to also know the conditional mean and variance of each y_j given $y_{j-1}, \dots, y_1$. This should remind us of forecasting! We didn't explicitly show $\mathbb{E}[y_j|y_{j-1}, \dots, y_1] = \hat{y}_j$ previously.

Rather, we derived $\hat{y}_j$ as the linear function of $y_{j-1}, \dots, y_1$ which minimizes

$$v_j = \mathbb{E}\left[\left(y_j - \hat{y}_j\right)^2 | y_{j-1}, \dots, y_1\right].$$

If we take the derivative with respect to $\hat{y}_j$, we obtain:

$$\mathbb{E}\left[2\left(y_j - \hat{y}_j\right) | y_{j-1}, \dots, y_1\right] = 0 \implies \mathbb{E}\left[y_j | y_{j-1}, \dots, y_1\right] = \hat{y}_j.$$

This confirms that $\mathbb{E}\left[y_j | y_{j-1}, \dots, y_1\right] = \hat{y}_j$, so we know the mean of each conditional distribution. Based on our definition of v_j this means that we also know that the variance of each conditional distribution is v_j.

Putting it all together, if $\boldsymbol{y}$ is distributed according to an $\mathbf{ARMA}(p, q)$ model we have:

- $y_j | y_{j-1}, \dots, y_1$ is normal;
- $\mathbb{E}\left[y_j | y_{j-1}, \dots, y_1\right] = \hat{y}_j$;
- $\mathbb{V}\left[y_j | y_{j-1}, \dots, y_1\right] = v_j$.

Then we can rewrite (11) as

$$p\left(\boldsymbol{y} | \boldsymbol{\phi}, \boldsymbol{\theta}, \mu, \sigma^2_w\right) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi v_j}} \exp\left\{-\frac{1}{2}\left(y_j - \hat{y}_j\right)^2 / v_j\right\}. \tag{12}$$

This is much nicer than (11) to work with - we saw when we discussed forecasting that there are multiple ways to quickly compute $\hat{y}_1, \dots, \hat{y}_n$ and $v_1, \dots, v_n$.

The next step is to take the log of (12), for computational stability, and maximize it over $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, μ, and σ^2_w. However, before we continue we are going to reparametrize (11) slightly to make this easier by replacing $v_j = \sigma^2_w r_j$. When we reparametrize in this way, r_j will not depend on σ^2_w.

$$p\left(\boldsymbol{y} | \boldsymbol{\phi}, \boldsymbol{\theta}, \mu, \sigma^2_w\right) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi \sigma^2_w r_j}} \exp\left\{-\frac{1}{2\sigma^2_w}\left(y_j - \hat{y}_j\right)^2 / r_j\right\}. \tag{13}$$

This allows us to split estimation of the $\mathbf{ARMA}(p, q)$ parameter σ^2_w look more like a least-squares problem.

Last, most maximum likelihood problems are nicer to work with on the log scale, so from here on out we'll work with two times the negative log-likelihood corresponding to (20),

$$n\log\left(\sigma_w^2\right) + \left(\sum_{j=1}^{n} \log\left(r_j\right)\right) + \left(\frac{1}{\sigma_w^2}\sum_{j=1}^{n} \frac{\left(y_j - \hat{y}_j\right)^2}{r_j}\right), \tag{14}$$

ignoring all constant terms that do not depend on the parameters of the **ARMA**(p, q) model. Now, we'll describe how three different maximum likelihood estimation procedures, unconditional maximum likelihood, unconditional least-squares, and conditional maximum likelihood.

Unconditional Maximum Likelihood: Unconditional maximum likelihood computes estimates of the parameters **ARMA**(p, q) model by maximizing (14) as written. First, notice that we can differentiate with respect to σ_w^2 to find that the optimal maximum likelihood maximizing σ_w^2 is given by

$$\sigma_w^2 = \frac{1}{n}\sum_{j=1}^{n} \frac{\left(y_j - \hat{y}_j\right)^2}{r_j}. \tag{15}$$

This means that we can plug this expression for σ_w^2 into (14), and just worry about finding the maximum likelihood estimates of $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and μ by minimizing:

$$n\log\left(\sum_{j=1}^{n} \frac{\left(y_j - \hat{y}_j\right)^2}{r_j}\right) + \left(\sum_{j=1}^{n} \log\left(r_j\right)\right). \tag{16}$$

Note that $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and μ enter into (16) via $\hat{y}_j$ and r_j. Maximizing (16) over $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and μ is a difficult nonlinear optimization problem. We're not going to talk about exactly how to solve it, but one important aspect to keep in mind is that it is necessary to require that we need to require that $\phi_1, \dots, \phi_p$ correspond to a stationary process. This is achieved by solving a reparametrized, linearly constrained optimization problem as described in Jones (1980), "Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations."

Once we've obtained estimates $\hat{\phi}_{UM,1}, \dots, \hat{\phi}_{UM,p}$, $\hat{\theta}_{UM,1}, \dots, \hat{\theta}_{UM,q}$, and $\hat{\mu}_{x,UM}$ from minimizing (16), we can recover $\hat{\sigma}^2_{w,UM}$ by plugging $\hat{y}_{UM,1}, \dots, \hat{y}_{UM,n}$ and $\hat{r}_{UM,1}, \dots, \hat{r}_{UM,n}$ into (15).

Unconditional Least Squares: Remember, we are using the reparameterized likelihood (20) because it more closely resembles a least-squares problem. There is one term that does not fit into the least-squares framework - $\left(\sum_{j=1}^n \log\left(r_j\right)\right)$. This term also tends to create extra nonlinearity that makes (14) especially challenging to solve.

Unconditional least-squares addresses this by eliminating this term altogether, and instead maximizing (21)

$$n\log\left(\sigma^2_w\right) + \left(\frac{1}{\sigma^2_w}\sum_{j=1}^n \frac{\left(y_j - \hat{y}_j\right)^2}{r_j}\right). \tag{17}$$

As with unconditional maximum likelihood, we can plug the least-squares σ^2_w, (21),

$$\sigma^2_w = \frac{1}{n-q-p-1}\sum_{j=1}^n \frac{\left(y_j - \hat{y}_j\right)^2}{r_j}. \tag{18}$$

into (14). This lets us ignore σ^2_w and focus on minimization over $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and μ. When we ignore the terms that do not depend on $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and μ and exponentiate again, this gives us

$$\sum_{j=1}^n \frac{\left(y_j - \hat{y}_j\right)^2}{r_j}. \tag{19}$$

Again, note that $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and μ enter into (19) via $\hat{y}_j$ and r_j. This can be easier to solve than (16), although it will still be a complicated nonlinear optimization problem. Again, we're not going to talk about exactly how to solve it. However, we note that we do need to explicitly constrain estimates of $\phi_1, \dots, \phi_p$ to be stationary. In the unconditional likelihood maximization, explicitly constraining $\phi_1, \dots, \phi_p$ to be stationary was not necessary because the term $\sum_{j=1}^n \log\left(r_j\right)$ would blow up to $+\infty$ if the constraint was violated.

Once we've obtained estimates $\hat{\phi}_{UL,1}, \dots, \hat{\phi}_{UL,p}$, $\hat{\theta}_{UL,1}, \dots, \hat{\theta}_{UL,q}$, and $\hat{\mu}_{x,UL}$ from minimizing (19), we can recover $\hat{\sigma}^2_{w,UL}$ by plugging $\hat{y}_{UL,1}, \dots, \hat{y}_{UL,n}$ and $\hat{r}_{UL,1}, \dots, \hat{r}_{UL,n}$ into (18).

Conditional Least Squares: There is one more way we can modify the maximum likelihood problem to make it even simpler. Let's go back to the likelihood (20). We could condition on the first $m = \max\{p, q\}$ values of the time series. Conveniently, when we condition on the first m values, r_j becomes constant because we can forecast any y_j equally well if we observed at least m previous values. Setting $r_j = r$ and conditioning on $y_1, \dots, y_m$, the conditional likelihood is

$$p\left(y_{m+1}, \dots, y_n | y_1, \dots, y_m \boldsymbol{\phi}, \boldsymbol{\theta}, \mu, \sigma^2_w\right) = \prod_{j=m+1}^{n} \frac{1}{\sqrt{2\pi\sigma^2_w r}} \exp\left\{-\frac{1}{2\sigma^2_w r}\left(y_j - \hat{y}_j\right)^2\right\}. \tag{20}$$

Dropping the $\frac{1}{\sqrt{r}}$ term that does not fit into the least-squares framework as well as the rest of the terms that do not depend on the **ARMA**(p, q) parameters, taking the log and multiplying by negative two, the conditional least-squares estimates of $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, μ, and σ^2_w will minimize:

$$(n - m)\log\left(\sigma^2_w\right) + \left(\frac{1}{\sigma^2_w r}\sum_{j=m+1}^{n}\left(y_j - \hat{y}_j\right)^2\right). \tag{21}$$

Just like we did with the unconditional maximum likelihood and unconditional least-squares problems, we can The least-squares σ^2_w is given by

$$\sigma^2_w = \frac{1}{r\left(n - m - q - p - 1\right)}\sum_{j=m+1}^{n}\left(y_j - \hat{y}_j\right)^2. \tag{22}$$

Plugging this into (21), dropping terms that do not depend on $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and μ and exponentiating yields a straightforward least-squares criterion that we can maximize over $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and μ,

$$\frac{1}{r}\sum_{j=m+1}^{n}\left(y_j - \hat{y}_j\right)^2. \tag{23}$$

Once more, note that $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and μ enter into (23) via $\hat{y}_j$ and r.

As in the unconditional least-squares case, we will need to explicitly constrain $\phi_1, \dots, \phi_p$

to be stationary because nothing in (23) prevents a minimum value being achieved at non-stationary $\phi_1, \dots, \phi_p$. Once again, for general $\mathbf{ARMA}(p, q)$ models (23) will not be linear in the $\mathbf{ARMA}(p, q)$ parameters $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$, and μ. As such, minimizing (23) will require nonlinear optimization methods, the details of which are beyond the scope of this class.

Once we've obtained estimates $\hat{\phi}_{CL,1}, \dots, \hat{\phi}_{CL,p}$, $\hat{\theta}_{CL,1}, \dots, \hat{\theta}_{CL,q}$, and $\hat{\mu}_{x,CL}$ from minimizing (23), we can recover $\hat{\sigma}^2_{w,CL}$ by plugging $\hat{y}_{CL,1}, \dots, \hat{y}_{CL,n}$ and $\hat{r}_{CL,1}, \dots, \hat{r}_{CL,n}$ into (22).

There is one special case where (23) **is linear** in the unknown parameters - the $\mathbf{AR}\left(p\right)$ model. In this case, $m = p$. When we were deriving the the Yule-Walker equations for the $\mathbf{AR}\left(p\right)$ model, we saw that the forecasts are given by $\hat{y}_j = \mu + \sum_{k=1}^p \phi_k \left(y_{j-k} - \mu\right)$ for $j > p$. Under the $\mathbf{AR}\left(p\right)$ model, it is easy to see that

$$\begin{aligned} r_j &= \mathbb{E}\left[\left(y_j - \hat{y}_j\right)^2 | y_{j-1}, \dots, y_{j-p}\right] / \sigma^2_w \\ &= \mathbb{E}\left[\left(y_j - \mu - \sum_{k=1}^p \phi_k \left(y_{j-k} - \mu\right)\right)^2 | y_{j-1}, \dots, y_{j-p}\right] / \sigma^2_w = 1. \end{aligned}$$

Then (23) simplifies to

$$\sum_{j=m+1}^n \left(y_j - \mu - \sum_{k=1}^p \phi_k \left(y_{j-k} - \mu\right)\right)^2. \tag{24}$$

The equation (24) is the same as the least-squares objective we would get if we regressed $\tilde{\boldsymbol{y}} = \left(y_{p+1}, \dots, y_n\right)$ on a intercept $\mathbf{1}_{n-p}$ and $(n-p) \times p$ design matrix $\boldsymbol{Z}$ with elements made up of lagged values $z_{ij} = y_{p+i-j}$. This means that conditional maximum likelihood estimation of ϕ_1, ..., ϕ_p under an $\mathbf{AR}\left(p\right)$ model can be performed using standard regression techniques!

Even better, it turns out that the conditional least-squares estimates $\hat{\phi}_{CL,1}, \dots, \hat{\phi}_{CL,p}$, $\hat{\mu}_{x,CL}$, and $\hat{\sigma}^2_{w,CL}$ are very similar to the Yule-Walker estimates $\hat{\phi}_{YW,1}, \dots, \hat{\phi}_{YW,p}$, $\hat{\mu}_{x,YW}$, and $\hat{\sigma}^2_{w,YW}$. This is easiest to see when we $x = 0$ and we assume that $\mu = 0$. Then

$$\hat{\boldsymbol{\phi}}_{CL} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\tilde{\boldsymbol{y}} = \left(\frac{1}{n-p}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \left(\frac{1}{n-p}\boldsymbol{Z}'\tilde{\boldsymbol{y}}\right). \tag{25}$$

The jk-th elements of $\frac{1}{n-p}\boldsymbol{Z}'\boldsymbol{Z}$ are given by $\frac{1}{n-p}\sum_{i=m+1}^{n} y_{i-j}y_{i-k}$, and the j-th element of $\frac{1}{n-p}\boldsymbol{Z}'\tilde{\boldsymbol{y}}$ is given by $\frac{1}{n-p}\sum_{i=m+1}^{n} y_{i-j}y_i$. These are sample autocovariances computed from $\tilde{\boldsymbol{y}}$! The only difference between $\hat{\boldsymbol{\phi}}_{CL}$ computed from (25) and $\hat{\boldsymbol{\phi}}_{YW}$ computed from (3) is whether or not the first p elements of $\boldsymbol{y}$ are used to compute the sample autocorrelations!

Asymptotic Distributions: Conveniently, all three of these maximum likelihood estimation procedures yield estimators with the same asymptotic distributions. This means that the behavior of our estimators won't depend much on whether we use unconditional maximum likelihood, unconditional least squares, or conditional least squares as long as n is large, i.e. we observe a long time series. Specifically, as $n \to \infty$

$$\sqrt{n}\left(\left(\begin{array}{c}\hat{\boldsymbol{\phi}}\\ \hat{\boldsymbol{\theta}}\end{array}\right)-\left(\begin{array}{c}\boldsymbol{\phi}\\ \boldsymbol{\theta}\end{array}\right)\right)\stackrel{d}{\to}\mathcal{N}\left(\mathbf{0},\left(\begin{array}{cc}\boldsymbol{\Sigma}_{\phi\phi} & \boldsymbol{\Sigma}_{\phi\theta}\\ \boldsymbol{\Sigma}_{\theta\phi} & \boldsymbol{\Sigma}_{\theta\theta}\end{array}\right)\right),$$

where $\boldsymbol{\Sigma}_{\phi\phi}$ is a $p\times p$ matrix, $\boldsymbol{\Sigma}_{\phi\theta}$ is a $p\times q$ matrix, $\boldsymbol{\Sigma}_{\theta\phi}$ is a $q\times p$ matrix, and $\boldsymbol{\Sigma}_{\theta\theta}$ is a $q\times q$ matrix. We can derive the elements of the covariance matrix $\boldsymbol{\Sigma}$ by introducing the $\mathbf{AR}\,(p)$ and $\mathbf{AR}\,(q)$ processes

$$\begin{aligned}
u_t &= \phi_1 u_{t-1} + \cdots + \phi_p u_{t-p} + w_t\\
v_t &= -\theta_1 v_{t-1} - \cdots - \theta_p v_{t-q} + w_t.
\end{aligned}$$

The elements of the covariance matrix are given by

$$\begin{aligned}
\sigma_{\phi\phi,ij} &= \mathbb{E}\left[u_t u_{t-(i-j)}\right]\\
\sigma_{\phi\theta,ij} &= \sigma_{\theta\phi,ji} = \mathbb{E}\left[u_t v_{t-(i-j)}\right]\\
\sigma_{\theta\theta,ij} &= \mathbb{E}\left[v_t v_{t-(i-j)}\right].
\end{aligned}$$

Model Selection

Moment-Based

If we are fitting an $\mathbf{AR}(p)$ or $\mathbf{MA}(q)$ model, we can actually figure out which coefficients to keep based on the sample partial-autocorrelations or the sample autocorrelations, respectively. Letting $v \sim \mathcal{N}(0,1)$, we previously learned that the sample autocorrelation is approximately normal $\hat{\gamma}_y(h) \approx v/\sqrt{n}$ under the null hypothesis that $\gamma_y(h) = 0$ as $n \to \infty$. Noting that $\gamma_y(h) = 0$ when $h > q$ if $\boldsymbol{y}$ is distributed according to a $\mathbf{MA}(q)$ model, this allows us to select the order of a $\mathbf{MA}$ model based on the sample autocorrelations.

Similarly,

$$\hat{c}_{jj} \approx v/\sqrt{n},\, v \sim \mathcal{N}(0,1) \tag{26}$$

when $j > p$ and $\boldsymbol{y}$ is distributed according to an $\mathbf{AR}(p)$ model as $n \to \infty$. This allows us to select the order of a $\mathbf{AR}$ model based on the sample partial autocorrelations.

It isn't quite so simple if we want to fit an $\mathbf{ARMA}(p,q)$ model. In that case, we'll want to consider maximum-likelihood based methods.

Likelihood-Based

Recall our definitions of AIC, AICc and SIC/BIC. We can compute select p and q by finding the values that minimize one of these quantities.

- $AIC = \ln\left(\hat{\sigma}^2_{w,UM}\right) + \frac{n+2(p+q+1)}{n}$.;
- $AICc = \ln\left(\hat{\sigma}^2_{w,UM}\right) + \frac{n+p+q+1}{n-p-q-1-2}$;
- $SIC = \ln\left(\hat{\sigma}^2_{w,UM}\right) + \frac{(p+q+1)\log(n)}{n}$.

Other

We won't discuss these methods in detail in class, but some other ways one might select the order of an $\mathbf{ARMA}(p, q)$ model include choosing the values of p and q:

- To minimize k-step-ahead forecasting error (kind of like leave-k-out cross validation);
- According to other diagnostics, e.g. smallest values of p and q that yield residuals that "appear" independent or fail to reject a null hypothesis of independence;
- Based on F-tests if using conditional likelihood to obtain estimates of $\boldsymbol{\phi}$ for $\mathbf{AR}(p)$ models, taking care to fit all models to the same $n - p_{max}$ observations, where p_{max} is the largest order being considered.