# STAT 804: Notes on Lecture 8

### Fitting $ARIMA(p, d, q)$ models to data

Fitting the $I$ part is easy we simply difference $d$ times. The same observation applies to seasonal multiplicative model. Thus to fit an ARIMA$(p, d, q)$ model to $X$ you compute $Y = (I - B)^d X$ (shortening your data set by $d$ observations) and then you fit an $ARMA(p, q)$ model to $Y$. So we assume that $d = 0$.

**Simplest case**: fitting the AR(1) model

$$X_t = \mu + \rho(X_{t-1} - \mu) + \epsilon_t$$

We must estimate 3 parameters: $\mu, \rho$ and $\sigma^2 = \text{Var}(\epsilon_t)$.

Our basic strategy will be:

- Estimate the parameters by maximum likelihood as if the series were Gaussian.
- Investigate the properties of the estimates for non-Gaussian data.

Generally the full likelihood is rather complicated; we will use conditional likelihoods and ad hoc estimates of some parameters to simplify the situation.

### The likelihood: Gaussian data

If the errors $\epsilon$ are normal then so is the series $X$. In general the vector $X = (X_0, \ldots, X_{T-1})^t$ has a $MVN(\mu, \Sigma)$ where $\Sigma_{ij} = C(i - j)$ and $\mu$ is a vector all of whose entries are $\mu$. The joint density of $X$ is

$$f_X(x) = \frac{1}{(2\pi)^{T/2} \det(\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)^t \Sigma^{-1}(x - \mu)\right\}$$

so that the log likelihood is

$$\ell(\mu, a_1, \ldots, a_p, b_1, \ldots, b_q, \sigma) = -\frac{1}{2}\left[(x - \mu)^t \Sigma^{-1}(x - \mu) + \log(\det(\Sigma))\right]$$

Here I have indicated precisely (for an $ARMA(p, q)$) the parameters on which the quantity depends.

It is possible to carry out full maximum likelihood by maximizing the quantity in question numerically. In general this is hard, however.

Here I indicate some standard tactics. In your homework I will be asking you to carry through this analysis for one particular model.

## The $AR(1)$ model

Consider the model

$$X_t - \mu = \rho(X_{t-1} - \mu) + \epsilon_t$$

This model formula permits us to write down the joint density of $X$ in a simpler way:

$$f_X = f_{X_{T-1}|X_{T-2},\ldots,X_0} f_{X_{T-2}|X_{T-3},\ldots,X_0} \cdots f_{X_1|X_0} f_{X_0}$$

Each of the conditional densities is simply

$$f_{X_{k+1}|X_k,\ldots,X_0}(x_k|x_{k-1},\ldots,x_0) = g\left[x_k - \mu - \rho(x_{k-1} - \mu)\right]$$

where $g$ is the density of an individual $\epsilon$. For iid $N(0,\sigma^2)$ errors this gives a log likelihood which is

$$\ell(\mu,\rho,\sigma) = -\frac{1}{2\sigma^2}\sum_1^{T-1}\left[x_k - \mu - \rho(x_{k-1} - \mu)\right]^2 - (T-1)\log(\sigma) + \log(f_{X_0})$$

Now for a stationary series I showed that $X_t \sim N(\mu, \sigma^2/(1-\rho^2))$ so that

$$\log(f_{X_0}(x_0)) = -\frac{1-\rho^2}{2\sigma^2}(x_0 - \mu)^2 - \log(\sigma) + \log(1-\rho^2)$$

This makes

$$\begin{aligned}\ell(\mu,\rho,\sigma) = &-\frac{1}{2\sigma^2}\left\{\sum_1^{T-1}\left[x_k - \mu - \rho(x_{k-1} - \mu)\right]^2 + (1-\rho^2)(x_0-\mu)^2\right\} \\ &- T\log(\sigma) + \log(1-\rho^2)\end{aligned}$$

We can maximize this over $\mu$ and $\sigma$ explicitly. First

$$\frac{\partial}{\partial\mu}\ell = \frac{1}{\sigma^2}\left\{\sum_1^{T-1}\left[x_k - \mu - \rho(x_{k-1} - \mu)\right](1-\rho) + (1-\rho^2)(x_0-\mu)\right\}$$

Set this equal to 0 to find

$$\begin{aligned}\hat\mu(\rho) &= \frac{(1-\rho)\sum_1^{T-1}(x_k - \rho x_{k-1}) + (1-\rho^2)x_0}{1-\rho^2+(1-\rho)^2(T-1)} \\ &= \frac{\sum_1^{T-1}(x_k - \rho x_{k-1}) + (1+\rho)x_0}{1+\rho+(1-\rho)(T-1)}\end{aligned}$$

Notice that this estimate is free of $\sigma$ and that if $T$ is large we may drop the 1 in the denominator and the term inolving $x_0$ in the denominator and get

$$\hat{\mu}(\rho) \approx \frac{\sum_1^{T-1}(x_k - \rho x_{k-1})}{(T-1)(1-\rho)}$$

Finally, the numerator is actually

$$\sum_0^{T-1} x_k - x_0 - \rho(\sum_0^{T-1} x_k - x_{T-1}) = (1-\rho)\sum_0^{T-1} x_k - x_0 + \rho x_{T-1}$$

The last two terms here are smaller than the sum; if we neglect them we get

$$\hat{\mu}(\rho) \approx \bar{X}\,.$$

Now compute

$$\frac{\partial}{\partial\sigma}\ell = \frac{1}{\sigma^3}\left\{\sum_1^{T-1}\left[x_k - \mu - \rho(x_{k-1}-\mu)\right]^2 + (1-\rho^2)(x_0-\mu)^2\right\} - \frac{T}{\sigma}$$

and set this to 0 to find

$$\hat{\sigma}^2(\rho) = \frac{\left\{\sum_1^{T-1}\left[x_k - \mu(\rho) - \rho(x_{k-1}-\mu(\rho))\right]^2 + (1-\rho^2)(x_0-\mu(\rho))^2\right\}}{T}$$

When $\rho$ is known it is easy to check that $(\mu(\rho), \sigma(\rho))$ maximizes $\ell(\mu, \rho, \sigma)$.

To find $\hat{\rho}$ you now plug $\hat{\mu}(\rho)$ and $\hat{\sigma}(\rho)$ into $\ell$ (getting the so called *profile likelihood* $\ell(\hat{\mu}(\rho), \rho, \hat{\sigma}(\rho))$) and maximize over $\rho$. Having thus found $\hat{\rho}$ the mles of $\mu$ and $\hat{\sigma}$ are simply $\hat{\mu}(\hat{\rho})$ and $\hat{\sigma}(\hat{\rho})$.

It is worth observing that fitted residuals can then be calculated:

$$\hat{\epsilon}_t = (X_t - \hat{\mu}) - \hat{\rho}(X_{t-1} - \hat{\mu})$$

(There are only $T-1$ of them since you cannot easily estimate $\epsilon_0$.) Note, too, that the formula for $\hat{\sigma}^2$ simplifies to

$$\hat{\sigma}^2 = \frac{\sum_1^{T-1}\hat{\epsilon}_t^2 + (1-\rho^2)(x_0-\mu(\rho))^2}{T} \approx \frac{\sum_1^{T-1}\hat{\epsilon}_t^2}{T}\,.$$

In general, we simplify the maximum likelihood problem several ways:

- We usually simply estimate $\hat{\mu} = \bar{X}$.

- The term $f_{X_0}$ in the likelihood is different in structure and causes considerable trouble. We drop it. The result is called a conditional likelihood. In general in a statistical inference problem if the data can be written in the form $X = (Y, Z)$ then we can factor the density in the form
$$f_X(x) = f_{Y|Z}(y|z) f_Z(z)$$
The first term in the factorization $f_{Y|Z}(y|z)$ is called a **conditional likelihood** (when you think of it as a function of the unknown parameters); the second term, $f_Z(z)$ is called a **marginal likelihood**. Sometimes one or the other of the two terms is conveniently simpler than the full likelihood; in these cases people often suggest using the simple piece. You get less efficient estimates in general but sometimes the loss is not very important.

  In the $AR(1)$ case $Y$ is just $(X_1, \ldots, X_{T-1})$ while $Z$ is $X_0$. We take our conditional log-likelihood to be
$$\begin{aligned}\ell(\mu, \rho, \sigma) &= \sum_1^{T-1} \log(f_{X_t|X_0,\ldots,X_{t-1}}) \\ &= \frac{-1}{2\sigma^2} \sum_1^{T-1} [X_t - \mu - \rho(X_{t-1} - \mu)]^2 - (T-1)\log(\sigma)\end{aligned}$$

- Combining the previous two ideas leads to maximization of
$$\ell(\bar{X}, \rho, \sigma) = \frac{-1}{2\sigma^2} \sum_1^{T-1} \left[X_t - \bar{X} - \rho(X_{t-1} - \bar{X})\right]^2 - (T-1)\log(\sigma)$$
This may be maximized explicitly to get
$$\hat{\rho} = \frac{\sum_1^{T-1}(X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_0^{T-2}(X_t - \bar{X})^2}$$
and
$$\hat{\sigma}^2 = \frac{\sum_1^{T-1}\left[X_t - \bar{X} - \hat{\rho}(X_{t-1} - \bar{X})\right]^2}{T-1}$$

- Changing the range of summation in the previously formula for $\hat{\rho}$ to include all possible terms gives

$$\hat{\rho} = \frac{\sum_1^{T-1}(X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_0^{T-1}(X_t - \bar{X})^2} = \frac{\hat{C}(1)}{\hat{C}(0)}$$

Notice that we have made a great many suggestions for simplifications and adjustments. This is typical of statistical research – many ideas, only slightly different from each other, are suggested and compared. In practice it seems likely that there is very little difference between all the methods. I am asking you in a homework problem to investigate the differences between several of these methods on a single data set.

## Higher order autoregressions

For the model

$$X_t - \mu = \sum_1^p a_i(X_{t-1} - \mu) + \epsilon_t$$

we will use conditional likelihood again. Let $\phi$ denote the vector $(a_1, \ldots, a_p)^t$. Now we condition on the first $p$ values of $X$ and use

$$\ell_c(\phi, \mu, \sigma) = -\frac{1}{2\sigma^2} \sum_p^{T-1} \left[ X_t - \mu - \sum_1^p a_i(X_{t-i} - \mu) \right]^2 - (T-p)\log(\sigma)$$

If we estimate $\mu$ using $\bar{X}$ we find that we are trying to maximize

$$-\frac{1}{2\sigma^2} \sum_p^{T-1} \left[ X_t - \bar{X} - \sum_1^p a_i(X_{t-i} - \bar{X}) \right]^2 - (T-p)\log(\sigma)$$

To estimate $a_1, \ldots, a_p$ then we merely minimize the sum of squares

$$\sum_p^{T-1} \hat{\epsilon}_t^2 = \sum_p^{T-1} \left[ X_t - \bar{X} - \sum_1^p a_i(X_{t-i} - \bar{X}) \right]^2$$

This is a straightforward regression problem. We regress the response vector

$$\begin{bmatrix} X_p - \bar{X} \\ \vdots \\ X_{T-1} - \bar{X} \end{bmatrix}$$

on the design matrix

$$\begin{bmatrix} X_{p-1}-\bar{X} & \cdots & X_0-\bar{X} \\ \vdots & \vdots & \vdots \\ X_{T-2}-\bar{X} & \cdots & X_{T-p-1}-\bar{X} \end{bmatrix}$$

An alternative to estimating $\mu$ by $\bar{X}$ is to define $\alpha = \mu(1-\sum a_i)$ and then recognize that

$$\ell(\alpha,\phi,\sigma) = -\frac{1}{2\sigma^2}\sum_p^{T-1}\left[X_t-\alpha-\sum_1^p a_i X_{t-i}\right]^2 - (T-p)\log(\sigma)$$

is maximized by regressing the vector

$$\begin{bmatrix} X_p \\ \vdots \\ X_{T-1} \end{bmatrix}$$

on the design matrix

$$\begin{bmatrix} X_{p-1} & \cdots & X_0 \\ \vdots & \vdots & \vdots \\ X_{T-2} & \cdots & X_{T-p-1} \end{bmatrix}$$

From $\hat{\alpha}$ and $\hat{\phi}$ we would get an estimate for $\mu$ by

$$\hat{\mu} = \frac{\hat{\alpha}}{1-\sum \hat{a}_i}$$

Notice that if we put

$$Z = \begin{bmatrix} X_{p-1}-\bar{X} & \cdots & X_0-\bar{X} \\ \vdots & \vdots & \vdots \\ X_{T-2}-\bar{X} & \cdots & X_{T-p-1}-\bar{X} \end{bmatrix}$$

then

$$Z^t Z \approx T\begin{bmatrix} \hat{C}(0) & \hat{C}(1) & \cdots & \\ \hat{C}(1) & \hat{C}(0) & \cdots & \cdots \\ \vdots & \cdots & \ddots & \cdots \\ \cdots & \cdots & \hat{C}(1) & \hat{C}(0) \end{bmatrix}$$

and if

$$Y = \begin{bmatrix} X_p - \bar{X} \\ \vdots \\ X_{T-1} - \bar{X} \end{bmatrix}$$

then

$$Z^t Y \approx T \begin{bmatrix} \hat{C}(1) \\ \vdots \\ \hat{C}(p) \end{bmatrix}$$

so that the normal equations (from least squares)

$$Z^t Z \phi = Z^T Y$$

are nearly the Yule-Walker equations again.

### Full maximum likelihood

To compute a full mle of $\theta = (\mu, \phi, \sigma)$ you generally begin by finding preliminary estimates $\hat{\theta}$ say by one of the conditional likelihood methods above and then iterate via Newton-Raphson or some other scheme for numerical maximization of the log-likelihood.

### Fitting $MA(q)$ models

Here we consider the model with known mean (generally this will mean we estimate $\hat{\mu} = \bar{X}$ and subtract the mean from all the observations):

$$X_t = \epsilon_t - b_1 \epsilon_{t-1} - \cdots - b_q \epsilon_{t-q}$$

In general $X$ has a $MVN(0, \Sigma)$ distribution and, letting $\psi$ denote the vector of $b_i$s we find

$$\ell(\psi, \sigma) = -\frac{1}{2}\left[\log(\det(\Sigma)) - X^T \Sigma^{-1} X\right]$$

Here $X$ denotes the column vector of all the data. As an example consider $q = 1$ so that

$$\Sigma = \begin{bmatrix} \sigma^2(1+b_1^2) & -b_1\sigma^2 & 0 & \cdots & \cdots \\ -b_1\sigma^2 & \sigma^2(1+b_1^2) & -b_1\sigma^2 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \cdots \\ 0 & \cdots & \cdots & -b_1\sigma^2 & \sigma^2(1+b_1^2) \end{bmatrix}$$

It is not so easy to work with the determinant and inverse of matrices like this. Instead we try to mimic the conditional inference approach above but with a twist; we now condition on something we haven't observed — $\epsilon_{-1}$.

Notice that

$$\begin{aligned}
X_0 &= \epsilon_0 - b\epsilon_{-1} \\
X_1 &= \epsilon_1 - b\epsilon_0 \\
&= \epsilon_1 - b(X_0 + \epsilon_{-1}) \\
X_2 &= \epsilon_2 - b\epsilon_1 \\
&= \epsilon_2 - b(X_1 + b(X_0 + \epsilon_{-1})) \\
&\vdots \\
X_{t-1} &= \epsilon_{T-1} - b(X_{T-2} + b(X_{T-3} + \cdots \epsilon_{-1}))
\end{aligned}$$

Now imagine that the data were actually

$$\epsilon_{-1}, X_0, \ldots, X_{T-1}$$

Then the same idea we used for an $AR(1)$ would give

$$\begin{aligned}
\ell(b,\sigma) &= \log(f(\epsilon_{-1},\sigma)) + \log(f(X_0,\ldots,X_{T-1}|\epsilon_{-1},b,\sigma) \\
&= \log(f(\epsilon_{-1},\sigma)) + \sum_0^{T-1} \log(f(X_t|X_{t-1},\ldots,X_0,\epsilon_{-1},b,\sigma)
\end{aligned}$$

The parameters are listed in the conditions in this formula merely to indicate which terms depend on which parameters. For Gaussian $\epsilon$s the terms in this likelihood are squares as usual (plus logarithms of $\sigma$) leading to

$$\begin{aligned}
\ell(b,\sigma) =& \frac{-\epsilon_{-1}}{2\sigma^2} - \log(\sigma) \\
& - \sum_0^{T-1} \left[\frac{1}{2}(X_t - \psi X_{t-1} - \psi^2 X_{t-2} - \cdots - \psi^{t+1}\epsilon_{-1})^2 + \log(\sigma)\right]
\end{aligned}$$

We will estimate the parameters by maximizing this function after getting rid of $\epsilon_{-1}$ somehow.

**Method A**: Put $\epsilon_{-1} = 0$ since 0 is the most probable value and maximize

$$-T\log(\sigma) - \frac{1}{2\sigma^2}\sum_0^{T-1}\left[X_t - \psi X_{t-1} - \psi^2 X_{t-2} - \cdots - \psi^t X_0\right]^2$$

Notice that for large $T$ the coefficients of $\epsilon_{-1}$ are close to 0 for most $t$ and the remaining few terms are negligible relatively to the total.

**Method B**: **Backcasting** is the process of guessing $\epsilon_{-1}$ on the basis of the data; we replace $\epsilon_{-1}$ in the log likelihood by

$$\mathrm{E}(\epsilon_{-1}|X_0, \ldots, X_{T-1}) .$$

The problem is that this quantity depends on $b$ and $\sigma$.

We will use the **EM algorithm** to solve this problem.