SF2943 Time Series Analysis: Lecture 8

April 28, 2022

In this lecture, we will introduce the definition of the partial autocorrelation function for a time series $\{X_t\}$. We will show how to use the partial autocorrelation function to select the order of an AR(p) process. The remaining part of this lecture will focus on model fitting with an ARMA(p, q) process.

1 Partial Autocorrelation Function

Given a stationary time series $\{X_t\}$, its correlation function $\rho(h) \doteq \text{Corr}(X_{t+h}, X_t)$ indicates the dependency between X_{t+h} and X_t. However, this dependency might come the dependency between X_{t+h-j} and X_t for $j = 1, \ldots, h-1$. In order to understand the absolute dependency between X_{t+h} and X_t, we have to "eliminate" the intermediate effects coming from $X_{t+1}, \ldots, X_{t+h-1}$. This gives us the partial autocorrelation function.

Definition 1.1 *The partial autocorrelation function (pacf) $\alpha(\cdot)$ of a stationary time series $\{X_t\}$ is defined by $\alpha(0) = 1$ and*

$$\alpha(h) = Corr(X_h - P(X_h|X_1, \ldots, X_{h-1}), X_0 - P(X_0|X_1, \ldots, X_{h-1})).$$

An equivalent way to define the pacf α is as follows.

Definition 1.2 *The partial autocorrelation function (pacf) $\alpha(\cdot)$ of a stationary time series $\{X_t\}$ is defined by $\alpha(0) = 1$ and $\alpha(h) = \phi_{hh}$ which is the last component of $\boldsymbol{\phi}_h = (\phi_{h1}, \phi_{h2}, \ldots, \phi_{hh})$ satisfying*

$$\Gamma_h \boldsymbol{\phi}_h = \boldsymbol{\gamma}_h \tag{1.1}$$

with $\Gamma_h = (\gamma(i-j))_{i,j=1}^h$ and $\boldsymbol{\gamma}_h = (\gamma(1), \gamma(2), \ldots, \gamma(h))'$.

Remark 1.3 *As we mentioned in previous lectures, the solution $\boldsymbol{\phi}_h = (\phi_{h1}, \phi_{h2}, \ldots, \phi_{hh})$ to (1.3) gives the coefficients of the best one-step linear predictor*

$$P_h X_{h+1} = \phi_{h1} X_h + \cdots + \phi_{hh} X_1.$$

Let us consider a concrete example and find the pacf.

Example 1.1 (AR(p) process) *Consider an AR(p) process*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t,$$

where $\{Z_t\} \sim WN(0, \sigma^2)$. *Assume that* $\phi(z) \neq 0$ *for* $|z| \leq 1$, *i.e.,* $\{X_t\}$ *is causal.*

For $h > p$, since $X_h = \phi_1 X_{h-1} + \cdots + \phi_p X_{h-p} + Z_h$, we find

$$\begin{aligned}
&P(X_h|X_1, \ldots, X_{h-1}) \\
&= P(\phi_1 X_{h-1} + \cdots + \phi_p X_{h-p} + Z_h|X_1, \ldots, X_{h-1}) \\
&= \phi_1 P(X_{h-1}|X_1, \ldots, X_{h-1}) + \cdots + \phi_p P(X_{h-p}|X_1, \ldots, X_{h-1}) + P(Z_h|X_1, \ldots, X_{h-1}) \\
&= \phi_1 X_{h-1} + \cdots + \phi_p X_{h-p},
\end{aligned}$$

where the last equation holds since Z_h is uncorrelated to $X_1, \ldots, X_{h-1}$ and $P(X_k|X_1, \ldots, X_{h-1}) = X_k$ for any $k = 1, \ldots, h-1$. Therefore,

$$\begin{aligned}
&X_h - P(X_h|X_1, \ldots, X_{h-1}) \\
&= (\phi_1 X_{h-1} + \cdots + \phi_p X_{h-p} + Z_h) - (\phi_1 X_{h-1} + \cdots + \phi_p X_{h-p}) = Z_h
\end{aligned}$$

and

$$\begin{aligned}
\alpha(h) &= \mathrm{Corr}(X_h - P(X_h|X_1, \ldots, X_{h-1}), X_0 - P(X_0|X_1, \ldots, X_{h-1})) \\
&= \mathrm{Corr}(Z_h, X_0 - P(X_0|X_1, \ldots, X_{h-1})) = 0.
\end{aligned}$$

The last equation holds since Z_h is uncorrelated to $X_0, X_1, \ldots, X_{h-1}$ and $P(X_0|X_1, \ldots, X_{h-1})$ is a linear combination of $X_1, \ldots, X_{h-1}$.

For $h = p$, using a similar argument we can find that

$$P(X_p|X_1, \ldots, X_{p-1}) = \phi_1 X_{p-1} + \cdots + \phi_{p-1} X_1 + \phi_p P(X_0|X_1, \ldots, X_{p-1})$$

and

$$X_p - P(X_p|X_1, \ldots, X_{p-1}) = \phi_p X_0 - \phi_p P(X_0|X_1, \ldots, X_{p-1}) + Z_p.$$

This implies

$$\begin{aligned}
&\mathrm{Cov}(X_p - P(X_p|X_1, \ldots, X_{p-1}), X_0 - P(X_0|X_1, \ldots, X_{p-1})) \\
&= \phi_p \mathrm{Var}(X_0 - P(X_0|X_1, \ldots, X_{p-1})) + \mathrm{Cov}(Z_p, X_0 - P(X_0|X_1, \ldots, X_{p-1})) \\
&= \phi_p \mathrm{Var}(X_0 - P(X_0|X_1, \ldots, X_{p-1})).
\end{aligned}$$

The last equation holds due to the fact that Z_p is uncorrelated to $X_0, X_1, \ldots, X_{p-1}$ and $P(X_0|X_1, \ldots, X_{p-1})$ is a linear combination of $X_1, \ldots, X_{p-1}$.

Thus,

$$\begin{aligned}
&\alpha(p) \\
&= \mathrm{Corr}(X_p - P(X_p|X_1, \ldots, X_{p-1}), X_0 - P(X_0|X_1, \ldots, X_{p-1})) \\
&= \frac{\mathrm{Cov}(X_p - P(X_p|X_1, \ldots, X_{p-1}), X_0 - P(X_0|X_1, \ldots, X_{p-1}))}{\sqrt{\mathrm{Var}(X_p - P(X_p|X_1, \ldots, X_{p-1}))\mathrm{Var}(X_0 - P(X_0|X_1, \ldots, X_{p-1}))}} \\
&= \frac{\phi_p \mathrm{Var}(X_0 - P(X_0|X_1, \ldots, X_{p-1}))}{\sqrt{\mathrm{Var}(X_0 - P(X_0|X_1, \ldots, X_{p-1}))\mathrm{Var}(X_0 - P(X_0|X_1, \ldots, X_{p-1}))}} \\
&= \phi_p,
\end{aligned}$$

where we use

$$\mathrm{Var}(X_0 - P(X_0|X_1,\dots,X_{p-1})) = \mathrm{Var}(X_p - P(X_p|X_1,\dots,X_{p-1})) \tag{1.2}$$

for the third equality. Equation (1.2) is true since $P(X_0|X_1,\dots,X_{p-1}) = P(X_p|X_1,\dots,X_{p-1})$.

Consequently,

$$\alpha(h) = \begin{cases} \phi_p, & \text{for } h = p \\ 0, & \text{for } h > p \end{cases}.$$

Remark 1.4 *The pacf $\alpha(h)$ for an AR(p) process become 0 for all $h > p$, and the acf $\rho(h)$ for an MA(q) process become 0 for all $h > q$.*

Remark 1.5 *Given observations $\{x_1,\dots,x_n\}$, and assume that we think an AR(p) model is suitable for fitting the data, then one way to figure the order p is to compute the sample pacf $\hat{\alpha}(h)$ by letting $\hat{\alpha}(0) = 1$ and $\hat{\alpha}(h) = \hat{\phi}_{hh}$ which is the last component of $\hat{\boldsymbol{\phi}}_h = (\hat{\phi}_{h1}, \hat{\phi}_{h2}, \dots, \hat{\phi}_{hh})$ satisfying*

$$\hat{\Gamma}_h \hat{\boldsymbol{\phi}}_h = \hat{\boldsymbol{\gamma}}_h \tag{1.3}$$

with $\hat{\Gamma}_h = (\hat{\gamma}(i-j))_{i,j=1}^h$ and $\hat{\boldsymbol{\gamma}}_h = (\hat{\gamma}(1), \hat{\gamma}(2), \dots, \hat{\gamma}(h))'$. Then as $\hat{\rho}(h)$ is approximated normal when n is large, $\hat{\alpha}(h)$ should be compatible with $N(0, 1/n)$ for $h > q$. In particular, $\hat{\alpha}(h)$ should fall between $\pm 1.96/\sqrt{n}$ with probability about 0.95. Thus, an estimator of p is

$$\min\left\{ m : |\hat{\alpha}(h)| < \frac{1.96}{\sqrt{n}} \text{ for all } h > m \right\}.$$

Similarly, suppose we think a MA(q) model is suitable, then one way to figure out the order q is to compute the sample acf $\hat{\rho}(h)$ and an estimator of q is

$$\min\left\{ m : |\hat{\rho}(h)| < \frac{1.96}{\sqrt{n}} \text{ for all } h > m \right\}.$$

2 Fitting ARMA(p, q) Model

Recall an ARMA(p, q) process $\{X_t\}$ (with mean zero) satisfies

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

with $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$. Hence, given observations $\{x_1,\dots,x_n\}$, assuming we remove the trend and seasonality, and we also rescale the data so that the sample mean is zero, then to pick an ARMA(p, q) model, we have to estimate not only $\boldsymbol{\phi} = (\phi_1,\dots,\phi_p)'$ and $\boldsymbol{\theta} = (\theta_1,\dots,\theta_q)'$, but also σ^2 as well as the order (p, q).

For simplicity, we first assume that the order (p, q) are given, and we focus on selection of $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and σ^2. In this case, there are four techniques (in the book), including Yule-Walker estimation, Burg's algorithm, Innovations algorithm, and Hannan-Rissanen algorithm. We will only introduce the first one due to the time constraint.

2.1 Yule-Walker estimation

Consider fitting a causal AR(p) model (with a given p)

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t.$$

By multiplying both sides of the equation with X_{t-j}, $j = 0, 1, \ldots, p$ and taking expectation, we find

$$\begin{cases} \Gamma_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p \\ \sigma^2 = \gamma(0) - \boldsymbol{\phi}' \boldsymbol{\gamma}_p \end{cases}, \tag{2.1}$$

where $\Gamma_p = (\gamma(i-j))_{i,j=1}^p$, $\boldsymbol{\gamma}_p = (\gamma(1), \ldots, \gamma(p))'$, and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$. The equations (2.1) are called Yule-Walker equations.

Remark 2.1 *Notice that because $\rho(h) = \gamma(h)/\gamma(0)$. The Yule-Walker equations* (2.1) *can be rewritten as*

$$\begin{cases} R_p \boldsymbol{\phi} = \boldsymbol{\rho}_p \\ \sigma^2 = \gamma(0)[1 - \boldsymbol{\phi}' \boldsymbol{\rho}_p] \end{cases}, \tag{2.2}$$

where $R_p = (\rho(i-j))_{i,j=1}^p$, $\boldsymbol{\rho}_p = (\rho(1), \ldots, \rho(p))'$, and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$.

From here we know that the parameters $\boldsymbol{\phi}$ and σ^2 for an AR(p) model satisfy the Yule-Walker equations (2.2). However, given observations $\{x_1, \ldots, x_n\}$, we do not know the acf ρ, but we can compute the sample acf $\hat{\rho}$, and use it as an approximation of ρ. Therefore, we have the following:

Sample Yule-Walker Equations: Estimate $\boldsymbol{\phi}$ and σ^2 for an AR(p) model by

$$\hat{\boldsymbol{\phi}} = (\hat{\phi}_1, \ldots, \hat{\phi}_p)' = \hat{R}_p^{-1} \hat{\boldsymbol{\rho}}_p$$

and

$$\hat{\sigma}^2 = \hat{\gamma}(0)[1 - \hat{\boldsymbol{\phi}}' \hat{\boldsymbol{\rho}}_p] = \hat{\gamma}(0) \left[1 - \hat{\boldsymbol{\rho}}_p' \hat{R}_p^{-1} \hat{\boldsymbol{\rho}}_p\right],$$

where $\hat{\boldsymbol{\rho}}_p = (\hat{\rho}(1), \ldots, \hat{\rho}(p))' = \hat{\boldsymbol{\gamma}}_p / \hat{\gamma}(0)$.

Remark 2.2 *If $\hat{\gamma}(0) > 0$, then $\hat{R}_m^{-1}$ is invertible for every $m = 1, 2, \ldots$.*

Remark 2.3 *For large n, the Yule-Walker estimators $\hat{\boldsymbol{\phi}}$ satisfies*

$$\hat{\boldsymbol{\phi}} \approx N\left(\boldsymbol{\phi}, \frac{\sigma^2}{n} \Gamma_p^{-1}\right).$$

2.2 Maximum likelihood estimation

Another way to identify estimators is via the maximum likelihood estimation. Therefore, the first question we need to ask is that given an ARMA(p,q) process $\{X_t\}$, what is the likelihood function or joint density of $(X_1,\ldots,X_n)$? Unfortunately, we do not know this kind of information in general. Nevertheless, suppose this ARMA(p,q) process $\{X_t\}$ is also a Gaussian process (and assume we know p and q), then one can find the likelihood function

$$L(\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2)=\frac{1}{\sqrt{(2\pi\sigma^2)^n r_0\cdots r_{n-1}}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\frac{\left(X_j-\hat{X}_j\right)^2}{r_{j-1}}\right\}, \tag{2.3}$$

where $\hat{X}_j$ is the best one-step linear predictor, i.e., $\hat{X}_j=P(X_j|X_1,\ldots,X_{j-1})$ and $r_0,\ldots,r_{n-1}$ are some numbers that we are able to evaluate.

Even if $\{X_t\}$ is not Gaussian, it still makes sense to regard (2.3) as a measure of goodness of fit of the model to the data, and to choose the parameters $\boldsymbol{\phi},\boldsymbol{\theta}$, and σ^2 that maximize (2.3). Since $\log(x)$ is a monotone increasing function for $x>0$, maximizing $L(\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2)$ is equivalent to maximizing $\log L(\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2)$. By differentiating $\log L(\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2)$ partially with respect to σ^2, we find that the maximum likelihood estimators $\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}}$, and $\hat{\sigma}^2$ satisfy the following equations:

Maximum Likelihood Estimators: $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\theta}}$ are the values of $\boldsymbol{\phi},\boldsymbol{\theta}$ that minimize

$$\ell(\boldsymbol{\phi},\boldsymbol{\theta})=\log\left(\frac{1}{n}S(\boldsymbol{\phi},\boldsymbol{\theta})\right)+\frac{1}{n}\sum_{j=1}^{n}r_{j-1}$$

and

$$\hat{\sigma}^2=\frac{1}{n}S(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}}),$$

where

$$S(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}})=\sum_{j=1}^{n}\frac{\left(X_j-\hat{X}_j\right)^2}{r_{j-1}}.$$

Remark 2.4 *We can not solve the minimization of $\ell(\boldsymbol{\phi},\boldsymbol{\theta})$ analytically or numerically. Instead, we need to use numerical approximation methods, such as gradient descent, to search for (and approximate) the minimizer. Such numerical approximation methods often rely on the initial condition in order to obtain a good performance. Namely, we have to identify $\boldsymbol{\phi},\boldsymbol{\theta}$ such that they are sort of "close" to the minimizer, and use these $\boldsymbol{\phi},\boldsymbol{\theta}$ as the initial condition for our search for the minimizer. Hence, the four techniques mentioned in the book, including Yule-Walker estimators, are still very useful in providing good enough initial conditions.*

2.3 Order selection for mixed models

So far we only considered estimation of $\boldsymbol{\phi},\boldsymbol{\theta}$ and σ^2 in the case when we know the order p and q. At the end of Section 1, we mentioned how to estimate p for an AR(p) model and estimate q

for a MA(q) model. But how can we estimate p and q simultaneously? For models with $p > 0$ and $q > 0$, sample acf $\hat{\rho}$ and sample pacf $\hat{\alpha}$ are difficult to recognize suitable p and q. It turns out that a systematic way for order selection is by using the AICC criterion.

AICC criterion: Find p and q by minimization of the AICC statistic

$$\text{AICC} = -2 \log L \left(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q, \frac{1}{n} S(\boldsymbol{\phi}, \boldsymbol{\theta}) \right) + 2 \frac{(p+q+1)n}{n-p-q-2},$$

where $\boldsymbol{\phi}_p = (\phi_1, \ldots, \phi_p)'$ and $\boldsymbol{\theta}_q = (\theta_1, \ldots, \theta_q)'$.

Remark 2.5 *1. This minimization provides estimates for $p, q, \boldsymbol{\phi}_p, \boldsymbol{\theta}_q$, and σ^2.*

2. We also need to use numerical approximation methods to approximate the minimizer.

3. The form of AICC comes from the Kullback-Leibler discrepancy (aka relative entropy) between the Gaussian density with $\boldsymbol{\phi}_p, \boldsymbol{\theta}_q$ and the Gaussian density with $\hat{\boldsymbol{\phi}}_p, \hat{\boldsymbol{\theta}}_q$, where $\hat{\boldsymbol{\phi}}_p, \hat{\boldsymbol{\theta}}_q$ are the maximum likelihood estimators of $\boldsymbol{\phi}_p, \boldsymbol{\theta}_q$. The main point is that Kullback-Leibler discrepancy is a measurement for the difference between two probability distributions, and AICC is some kind of Kullback-Leibler discrepancy.

4. In general, given observations $\{x_1, \ldots, x_n\}$, it is possible to find p, q such that the ARMA(p, q) process $\{X_t\}$ fits the data perfectly, in the sense that, $\gamma_X(h) = \hat{\gamma}(h)$ for all $h = 1, \ldots, n$, where $\gamma_X(h)$ is the acvf for $\{X_t\}$ and $\hat{\gamma}(h)$ is the sample acvf computed by $\{x_1, \ldots, x_n\}$. However, the resulting p, q might be quite large, which cause the problem of over-fitting.

5. The first term in AICC corresponds to the log likelihood function, and the second term plays the role of a penalty for large values p, q. This penalty prevents over-fitting.