Spring 2013 Statistics 153 (Time Series) : Lecture Ten

Aditya Guntuboyina

21 February 2013

1 Best Linear Prediction

Suppose that Y and $W_1, \ldots, W_m$ are random variables with zero means and finite variances. Let $\text{cov}(Y, W_i) = \zeta_i, i = 1, \ldots, m$ and

$$\text{cov}(W_i, W_j) = \Delta(i, j) \qquad \text{for } i, j = 1, \ldots, m.$$

What is the best **linear predictor** of Y in terms of $W_1, \ldots, W_m$?

The best linear predictor $a_1, \ldots, a_m$ is characterized by the property that $Y - a_1 W_1 - \cdots - a_m W_m$ is uncorrelated with $W_1, \ldots, W_m$. In other words:

$$\text{cov}(Y - a_1 W_1 - \cdots - a_m W_m, W_i) = 0 \qquad \text{for } i = 1, \ldots, m.$$

Note that this gives m equations in the m unknowns $a_1, \ldots, a_m$. The ith equation can be rewritten as

$$\zeta_i - \Delta(i, 1) a_1 - \cdots - \Delta(i, m) a_m = 0.$$

In other words, this means that ζ_i equals the ith row of Δ multiplied by the vector $a = (a_1, \ldots, a_m)^T$ which is same as the ith element of the vector Δa. Thus these m equations can be written in one line as $\Delta a = \zeta$.

Another way to get this defining equation for the coefficients of the best linear predictor is to find values of $a_1, \ldots, a_m$ that minimize

$$\begin{aligned} F(\mathbf{a}) &:= \mathbb{E}\left(Y - a_1 W_1 - \cdots - a_m W_m\right)^2 \\ &= \mathbb{E}\left(Y - a^T W\right)^2 \\ &= \mathbb{E} Y^2 - 2\mathbb{E}((a^T W) Y) + \mathbb{E}(a^T W W^T a) \\ &= \mathbb{E} Y^2 - 2 a^T \zeta + a^T \Delta a. \end{aligned}$$

Differentiate with respect to a and set equal to zero to get

$$-2\zeta + 2\Delta a = 0$$

or $a = \Delta^{-1}\zeta$. Therefore the best linear predictor of Y in terms of $W_1, \ldots, W_m$ equals $\zeta^T \Delta^{-1} W$.

The special case of this for $m = 1$ (when there is only one predictor W_1) may be more familiar. When $m = 1$, we have $\zeta_1 = \text{cov}(Y, W_1)$ and $\Delta(1, 1) = \text{var}(W_1)$. Thus, the best predictor or Y in terms of W_1 is

$$\frac{\text{cov}(Y, W_1)}{\text{var}(W_1)} W_1.$$

Now consider a stationary mean zero time series $\{X_t\}$. Using the above with $Y = X_n$ and $W_1 = X_{n-1}$, we get that the best predictor of X_n in terms of X_{n-1} is

$$\frac{\text{cov}(X_n, X_{n-1})}{\text{var}(X_{n-1})} X_{n-1} = \frac{\gamma_X(1)}{\gamma_X(0)} X_{n-1} = \rho_X(1) X_{n-1}$$

What is the best predictor for X_n in terms of $X_{n-1}, X_{n-2}, \ldots, X_{n-k}$? Here we take $Y = X_n$ and $W_i = X_{n-i}$ for $i = 1, \ldots, k$. Therefore

$$\Delta(i,j) = \text{cov}(W_i, W_j) = \text{cov}(X_{n-i}, X_{n-j}) = \gamma_X(i-j)$$

and

$$\zeta_i = \text{cov}(Y, W_i) = \text{cov}(X_n, X_{n-i}) = \gamma_X(i).$$

With these Δ and ζ, solve for $\Delta a = \zeta$ to obtain the coefficients of $X_{n-1}, \ldots, X_{n-k}$ in the best linear predictor of X_n.

Consider the special case of the AR(p) model: $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$. Directly from the defining equation and causality, it follows that $X_n - \phi_1 X_{n-1} - \cdots - \phi_p X_{n-p}$ is uncorrelated with $X_{n-1}, X_{n-2}, \ldots$. We thus deduce that the best linear predictor of X_n in terms of $X_{n-1}, X_{n-2}, \ldots$ equals $\phi_1 X_{n-1} + \phi_2 X_{n-2} + \cdots + \phi_p X_{n-p}$.

2 The Partial Autocorrelation Function (pacf)

2.1 First Definition

Let $\{X_t\}$ be a mean zero stationary process. The Partial Autocorrelation at lag h, denoted by $pacf(h)$ is defined as the coefficient of X_{t-h} in the best linear predictor for X_t in terms of $X_{t-1}, \ldots, X_{t-h}$.

Check that $pacf(1)$ is the same as the autocorrelation at lag one, $\rho(1)$. But $pacf(h)$ for $h > 1$ can be quite different from $\rho(h)$.

For the AR(p) model: $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$, check that $pacf(p) = \phi_p$ and that $pacf(h) = 0$ for $h > p$.

2.2 Second Definition

From the first definition, it is not quite clear why this is called a correlation. This will be apparent from the second definition.

The pacf at lag h is defined as the correlation between X_t and X_{t-h} **with the effect of the intervening variables $X_{t-1}, X_{t-2}, \ldots, X_{t-h+1}$ removed**. Let $\beta_1 X_{t-1} + \cdots + \beta_{h-1} X_{t-h+1}$ denote the best linear predictor of X_t in terms of $X_{t-1}, \ldots, X_{t-h+1}$. By stationarity, the two sequences

$$X_t, X_{t-1}, \ldots, X_{t-h+1}$$

and

$$X_{t-h}, X_{t-h+1}, \ldots, X_{t-1}$$

have the same covariance matrix. Indeed, if $W_i = X_{t-i+1}$ and $\tilde{W}_i = X_{t-h+i-1}$ for $i = 1, \ldots, h$, then the covariance between W_i and W_j equals $\gamma_X(i-j)$ which is the same as the covariance between $\tilde{W}_i$ and $\tilde{W}_j$.

Therefore, the best linear prediction of X_{t-h} in terms of $X_{t-h+1}, \ldots, X_{t-1}$ equals $\beta_1 X_{t-h+1} + \cdots + \beta_{h-1} X_{t-1}$.

The pacf at lag h is defined as

$$pacf(h) = \text{corr}\left(X_t - \beta_1 X_{t-1} - \cdots - \beta_{h-1} X_{t-h+1}, X_{t-h} - \beta_1 X_{t-h+1} - \cdots - \beta_{h-1} X_{t-1}\right).$$

In other words, $pacf(h)$ is the correlation between the **errors in the best linear predictions** of X_t and X_{t-h} in terms of the intervening variables $X_{t-1}, \ldots, X_{t-h+1}$.

The key fact is that for an AR(p) model, $pacf(h)$ equals zero for lags $h > p$. To see this: note that for $h > p$, the best linear predictor for X_t in terms of $X_{t-1}, \ldots, X_{t-h+1}$ equals $\phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p}$. In other words, $\beta_1 = \phi_1, \ldots, \beta_p = \phi_p$ and $\beta_i = 0$ for $i > p$.

Therefore for $h > p$, we have

$$\begin{aligned} pacf(h) &= \text{corr}\left(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}, X_{t-h} - \phi_1 X_{t-h+1} - \cdots - \phi_p X_{t-h+p}\right) \\ &= \text{corr}\left(Z_t, X_{t-h} - \phi_1 X_{t-h+1} - \cdots - \phi_p X_{t-h+p}\right) = 0, \end{aligned}$$

by causality.

The equivalence between the two definitions of $pacf(h)$ can be proved by linear algebra. We will skip this derivation.

3 Estimating *pacf* from Data

How does one estimate $pacf(h)$ from data for different lags h? The coefficients $a_1, \ldots, a_h$ of $X_{t-1}, \ldots, X_{t-h}$ in the best linear predictor of X_t are obtained by solving an equation of the form $\Delta a = \zeta$.

Now all the elements of Δ and ζ are of the form $\gamma_X(i-j)$ for some i and j. Therefore, a natural method of estimating $pacf(h)$ is to estimate the entries in Δ and ζ by the respective sample autocorrelations to obtain $\hat{\Delta}$ and $\hat{\zeta}$ and then to solve the equation $\hat{\Delta}\hat{a} = \hat{\zeta}$ for $\hat{a}$. Note that $pacf(h)$ is precisely a_h.

It has been shown that when the data come from an AR(p) model, the sample partial autocorrelations at lags greater than p are approximately **independently normally** distributed with zero means and variances $1/n$. Thus for $h > p$, bands at $\pm 1.96 n^{-1/2}$ can be used for checking if an AR(p) model is appropriate.

4 Summary

For an $MA(q)$ model, the autocorrelation function $\rho_X(h)$ equals zero for $h > q$. Also for $h > q$, the sample autocorrelation functions r_h are approximately normal with mean 0 and variance w_{hh}/n where $w_{hh} := 1 + 2\rho^2(1) + \cdots + 2\rho^2(q)$.

For an $AR(p)$ model, the partial autocorrelation function $pacf(h)$ equals zero for $h > p$. Also for $h > p$, the sample autocorrelation functions r_h are approximately normal with mean 0 and variance $1/n$.

If the sample acf for a data set cuts off at some lag, we use an MA model. If the sample pacf cuts off at some lag, we use an AR model.

What if neither of the above happens? How do we then choose an appropriate ARMA model? Here is a general strategy:

1. Try ARMA(p, q) for various choices of p and q.
2. For a fixed p and q, fit the ARMA(p, q) model to the data (we will soon learn how to do this).

3. See how good the fit is. Select p and q so that the fit is good **while making sure there is no overfitting**.

How to check if a model is fits the data well but does not overfit? This is a problem of model selection. Often automatic criteria like AIC, FPE, BIC are used. One should also use judgement.

Our plan is as follows:

1. How to fit an ARMA model to data?
2. How to assess goodness of fit?
3. Choosing p and q by an automatic Model selection technique.