Course Notes for
Time Series

Fei Li*

*Email: fei.li.best@gmail.com.

Part I

Univariate Time Series Analysis

1 Introduction

Definition 1.1 (Autocovariance Function)**.** Let $\{X_t\}$ be a stochastic process. The *autocovariance function* is

$$\gamma_X(t,s) = \text{cov}(X_t, X_s) = \mathbb{E}X_t X_s - \mathbb{E}X_t \mathbb{E}X_s.$$

Definition 1.2 (Stationarity)**.** A stochastic process $\{X_t\}$ is called *stationary* if for all integers r, s and t

1. $\mathbb{E}X_t = \mu$ is constant;
2. $\mathbb{V}X_t < \infty$;
3. $\gamma_X(t,s) = \gamma_X(t+r, s+r)$.

The *autocorrelation function* (ACF) is defined as

$$\rho_X(h) = \frac{\rho_X(h)}{\rho_X(0)} = \text{corr}(X_{t+h}, X_t)$$

for all $h \in \mathbb{Z}$.

1.1 Construction of Stochastic Processes

The simplest building block is a process with zero autocorrelation called a white noise process.

1.1.1 White Noise

A stationary process $\{Z_t\}$ is called *white noise* if

- $\mathbb{E}Z_t = 0$;
- $\gamma_Z(h) = \begin{cases} \sigma^2 & h = 0; \\ 0 & h \neq 0 \end{cases}$

We denote this by $Z_t \sim WN(0, \sigma^2)$. The white noise process is therefore stationary and temporally uncorrelated, i.e. the ACF is always equal to zero, except for $h \neq 0$ where it is equal to one. As the ACF possesses no structure, it is impossible to draw inferences from past observations to its future development, at least in a least square setting with linear forecasting functions. Therefore one can say that a white noise process has no memory. If $\{Z_t\}$ is not only temporally uncorrelated, but also i.i.d. then we write $Z_t \sim IID(0, \sigma^2)$. If in addition Z_t is normally distributed, then we write $Z_t \sim IIN(0, \sigma^2)$.

1.1.2 Moving-Average of Order One

The moving-average process of order one, MA(1) process, is

$$X_t = Z_t + \theta Z_{t-1} \quad \text{with } Z_t \sim WN(0, \sigma^2).$$

We have $\mathbb{E}X_t = 0$ and

$$\begin{aligned}
\gamma_X(t+h, t) &= \operatorname{cov}(X_{t+h}, X_t) \\
&= \operatorname{cov}(Z_{t+h} + \theta Z_{t+h-1}, Z_t + \theta Z_{t-1}) \\
&= \mathbb{E}Z_{t+h}Z_t + \theta \mathbb{E}Z_{t+h}Z_{t-1} + \theta \mathbb{E}Z_{t+h-1}Z_t + \theta^2 \mathbb{E}Z_{t+h-1}Z_{t-1} \\
&= \begin{cases} (1+\theta^2)\sigma^2 & h = 0; \\ \theta\sigma^2 & h = \pm 1; \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

The autocorrelation function (ACF) is

$$\rho_X(h) = \begin{cases} 1 & h = 0; \\ \frac{\theta}{1+\theta^2} & h = \pm 1; \\ 0 & \text{otherwise.} \end{cases}$$

Note that the correlation between X_t and X_{t-1} is restricted within $\frac{1}{2}$.

1.1.3 Random Walk

Let $\{Z_t\}$ be a white noise process. Then the process

$$X_t = Z_1 + Z_2 + \cdots + Z_t = \sum_{j=1}^{t} Z_j, \quad t > 0$$

is called a *random walk*. In contrast to $\{Z_t\}$, $\{X_t\}$ is only defined for $t > 0$. The random walk can alternatively be defined through the recursion

$$X_t = X_{t-1} + Z_t, \quad t > 0 \text{ and } X_0 = 0.$$

If in each time period a constant δ is added such that $X_t = \delta + X_{t-1} + Z_t$, then the process is called a *random walk with drift*. The random walk process is not stationary, as $\mathbb{V}(X_t) = t \cdot \sigma^2$ depends on t.

2 ARMA Models

Definition 2.1. A stochastic process $\{X_t\}$ is called an *autoregressive moving-average process* (*ARMA* process) of order (p, q) if it is stationary and satisfies a linear stochastic difference equation of the form

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

with $Z_t \sim WN(0, \sigma^2)$ and $\phi_q \theta_q \neq 0$. $\{X_t\}$ is called an ARMA(p, q) process with mean μ of $\{X - \mu\}$ is an ARMA(p, q) process.

2.1 The Lag Operator

We use L to denote the Lag operator:

$$\mathrm{L} X_t = X_{t-1}.$$

We use the following to denote the autoregressive and moving-average polynomials:

$$\begin{aligned} \Phi(\mathrm{L}) &= 1 - \phi_1 \mathrm{L} - \cdots - \phi_p \mathrm{L}^p, \\ \Theta(\mathrm{L}) &= 1 + \theta_1 \mathrm{L} + \cdots + \theta_p \mathrm{L}^p. \end{aligned}$$

We write ARMA process as $\Phi(\mathrm{L}) X_t = \Theta(\mathrm{L}) Z_t$.

2.2 Some Special Cases

2.2.1 MA(q) Process

The MA(q) process is

$$Z_t = \Theta(\mathrm{L}) Z_t = \theta_0 Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \quad \text{with } \theta_0 = 1 \text{ and } \theta_q \neq 0.$$

First and second moments: $\mathbb{E} X_t = 0$ and

$$\begin{aligned} \gamma_X(h) &= \operatorname{cov}(X_{t+h}, X_t) = \mathbb{E}(X_{t+h} X_t) \\ &= \begin{cases} \sigma^2 \sum_{i=0}^{q-|h|} \theta_i \theta_{i+|h|}, & |h| \leq q; \\ 0 & |h| > q. \end{cases} \end{aligned}$$

Note that every MA(q) process is stationary irrespective of the parameters $\theta_1, \ldots, \theta_q$. Because the correlation between X_t and X_s is zero if t and s are more than q periods apart, we call such processes have *short memory* or *short range dependence*.

2.2.2 AR(1) Process

The AR(1) process is

$$X_t = \phi X_{t-1} + Z_t \quad Z_t \sim WN(0, \sigma^2) \text{ and } \phi \neq 0.$$

If $|\phi| < 1$ then substitute the formula for X_t several times into the difference equation we get

$$\begin{aligned} X_t &= \phi(\phi X_{t-2} + Z_{t-1}) + Z_t = \cdots \\ &= Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \cdots + \phi^k Z_{t-k} + \phi^{k+1} X_{t-k-1}. \end{aligned}$$

If $\{X_t\}$ is stationary, then $\mathbb{V} X_{t-k-1}$ is constant and does not depend on k, so we have

$$\mathbb{V}\left(X_t - \sum_{j=0}^{k} \phi^j Z_{t-j} \right) = \phi^{2k+2} \mathbb{V} X_{t-k-1} \to 0 \text{ as } k \to \infty.$$

This shows that $\sum_{j=0}^{k} \phi_j Z_{t-j}$ converges in the mean square sense, and thus also in probability, to X_t as $k \to \infty$. Thus we can take

$$X_t = Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \cdots = \sum_{j=0}^{\infty} \phi_j Z_{t-j}$$

as a solution to the stochastic difference equation. The mean and autocovariance function are $\mathbb{E} X_t = 0$ and

$$\gamma_X(h) = \operatorname{cov}(X_{t+h}, X_t) = \mathbb{E}\left(\sum_{j=0}^{\infty} \phi^j Z_{t+h-j} \right) \left(\sum_{j=0}^{\infty} \phi^j Z_{t-j} \right). \tag{1}$$

We match the terms in the first parenthesis and in the second parenthesis with equal subscripts. Let $t + h - j_1 = t - j_2$, we get $j_1 = h + j_2$. Thus the coefficients ϕ^{h+j_2} from the left and ϕ^{j_2} from the right will match. We have $\phi^{|h|} \cdot \phi^{2j_2} \cdot \sigma^2$ for every j_2, so $\gamma_X(h)$ is

$$\gamma_X(h) = \sigma^2 \phi^{|h|} \sum_{j=0}^{\infty} \phi^{2j} = \frac{\phi^{|h|}}{1 - \phi^2} \sigma^2, \quad h \in \mathbb{Z}.$$

The autocorrelation function is $\rho_X(h) = \gamma_X(h)/\gamma_X(0) = \phi^{|h|}$.

2.3 Causality and Invertibility

Definition 2.2. An ARMA(p, q) process $\{X_t\}$ with $\Phi(\mathrm{L}) X_t = \Theta(\mathrm{L}) Z_t$ is called *causal* with respect to $\{Z_t\}$ if there exists a sequence $\{\psi_j\}$ with $\sum_{j=0}^{\infty} |\psi_j| < \infty$ such that

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} = \Psi(\mathrm{L}) Z_t \quad \text{with } \psi_0 = 1$$

where $\Psi(\mathrm{L}) = \sum_{j=0}^{\infty} \psi_j L^j$.

As we can observe, the coefficients $\{\psi_j\}$ is exactly the impulse response function.

Theorem 2.3. Let $\{X_t\}$ be an ARMA(p,q) process with $\Phi(\mathrm{L})X_t = \Theta(\mathrm{L})Z_t$ and assume that the polynomials $\Phi(z)$ and $\Theta(z)$ have no common root. Then $\{X_t\}$ is causal with respect to $\{Z_t\}$ if and only if $\Phi(z) \neq 0$ for $|z| \leq 1$, i.e. all roots of the equation $\Phi(z) = 0$ lie outside the unit circle.

How do we calculate $\{\psi_j\}$? We expand $\Psi(z)\Phi(z) = \Theta(z)$, namely

$$(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots)(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$$

to get

$$\begin{array}{llll} \psi_0 & -\psi_0\phi_1 z & -\psi_0\phi_2 z^2 - \cdots - \psi_0\phi_p z^p + \\ & \psi_1 z & -\psi_1\phi_1 z^2 - \cdots - \psi_1\phi_p z^{p+1} + \\ & & \phi_2 z^2 - \cdots \end{array}$$

$= \psi_0 + (\psi_1 - \psi_0\phi_1)z + (\psi_2 - \psi_0\phi_2 - \psi_1\phi_1)z^2 + \cdots$. We equate the coefficients from $1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$ to get

$$\begin{aligned} \psi_0 &= 1, \\ \psi_1 &= \theta_1 + \phi_1\psi_0 = \theta_1 + \phi_1, \\ \psi_2 &= \theta_2 + \phi_2\psi_0 + \phi_1\psi_1 = \theta_2 + \phi_2 + \phi_1\theta_1 + \phi_1^2. \\ &\vdots \end{aligned}$$

As an exercise, given the causal representation of an ARMA$(2,1)$ process, how do you recover the coefficients $\{\phi_1, \phi_2, \theta_1\}$? The equation is

$$(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots)(1 - \phi_1 z - \phi_2 z^2) = 1 + \theta_1 z.$$

We have three unknowns so we need three equations. Multiply out:

$$\begin{cases} \theta_1 + \phi_1 = \psi_1 \\ \psi_1\phi_1 + \psi_0\phi_2 = \psi_2 \\ \psi_2\phi_1 + \psi_1\phi_2 = \psi_3 \end{cases}$$

And then we can solve for $\{\phi_1, \phi_2, \theta_1\}$.

In time series analysis the realizations of $\{X_t\}$ are observed and $\{Z_t\}$ are unobserved. It is therefore of interest to know whether it is possible to recover the unobserved shocks from the observations $\{X_t\}$. This leads to the concept of invertibility.

Definition 2.4. An ARMA(p,q) process for $\{X_t\}$ with $\Phi(\mathrm{L})X_t = \Theta(\mathrm{L})Z_t$ is called *invertible* with respect to $\{Z_t\}$ if there exists a sequence $\{\pi_j\}$ with $\sum_{j=0}^{\infty} |\pi_j| < \infty$ such that

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

Theorem 2.5. $\{X_t\}$ is invertible with respect to $\{Z_t\}$ if and only if $\Theta(z) \neq 0$ for $|z| \leq 1$.

2.4 Computation of the Autocovariance Function of an ARMA process

How to compute the autocovariance function $\gamma_X(h)$ of an ARMA process $\{X_t\}$?

2.4.1 First Procedure

We can write $\{X_t\}$ in terms of its causal representation $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, and then calculate $\gamma_X(h)$ as in Eq. (1):

$$\gamma_X(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}.$$

2.4.2 Second Procedure

We can multiply the ARMA equation by X_{t-h} and apply the expectation operator:

$$\begin{aligned}\mathbb{E}X_t X_{t-h} - \phi_1 \mathbb{E}X_{t-1}X_{t-h} - \cdots - \phi_p \mathbb{E}X_{t-p}X_{t-h} \\ = \mathbb{E}Z_t X_{t-h} + \theta_1 \mathbb{E}Z_{t-1}X_{t-h} + \cdots + \theta_q \mathbb{E}Z_{t-q}X_{t-h}.\end{aligned}$$

This leads to an equation system

$$\begin{aligned}\gamma(h) - \phi_1\gamma(h-1) - \cdots - \phi_p\gamma(h-p) &= \sigma^2 \textstyle\sum_{h \le j \le q} \theta_j \psi_{j-h}, \quad && h < \max\{p, q+1\} \\ \gamma(h) - \phi_1\gamma(h-1) - \cdots - \phi_p\gamma(h-p) &= 0, && h \ge \max\{p, q+1\}.\end{aligned}$$

The general solution of the difference equation is

$$\gamma(h) = c_1 z_1^{-h} + \cdots + c_p z_p^{-h}$$

where $z_1, \ldots, z_p$ are the distinct roots of the polynomial $\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = 0$. Note that the initial conditions $\psi_1, \ldots, \psi_q$ have to be determined beforehand.

3 Forecasting Stationary Process

The problem is: given the observations $\{X_T, \ldots, X_1\}$, how do we forecast X_{T+h} for $h > 0$?

3.1 The Theory of Linear Least-Squares Forecasts

We restrict ourselves to *linear* forecast functions, also called *linear predictors*, $\mathbb{P}_T X_{T+h}$. It takes the form

$$\mathbb{P}_T X_{T+h} = a_0 + a_1 X_T + \cdots + a_T X_1 = a_0 + \sum_{i=1}^{T} a_i X_{T+1-i}.$$

We determine the coefficients by minimizing the mean squared errors:

$$\min_{a_0,\ldots,a_T} S = S(a_0, \ldots, a_T) = \mathbb{E}(X_{T+h} - \mathbb{P}X_{T+h})^2 = \mathbb{E}(X_{T+h} - a_0 - a_1 X_T - \cdots - a_T X_1)^2.$$

The first order conditions are

$$\frac{\partial S}{\partial a_0} = \mathbb{E}\left(X_{T+h} - a_0 - \sum_{i=1}^{T} a_i X_{T+1-i}\right) = 0, \tag{2}$$

$$\frac{\partial S}{\partial a_j} = \mathbb{E}\left[\left(X_{T+h} - a_0 - \sum_{i=1}^{T} a_i X_{T+1-i}\right) X_{T+1-j}\right] = 0, \quad j = 1, \ldots, T. \tag{3}$$

The first equation says that the mean of the forecast error $\mathbb{E}(X_{T+h} - \mathbb{P}X_{T+h})$ is zero. There is no bias, neither upward nor downward, in the forecasts. The second equation says that $\mathbb{E}[(X_{T+h} - \mathbb{P}_T X_{T+h})X_{T+1-j}] = 0$ for $j = 1, \ldots, T$. The forecast error is orthogonal (uncorrelated) with the available information represented by the past observations. Geometrically speaking, the best linear forecast is obtained by finding the point in the linear subspace spanned by $\{X_T, \ldots, X_1\}$ which is closest to X_{T+h}. The normal equations can be rewritten in matrix notation as

$$a_0 = \mu\left(1 - \sum_{i=1}^{T} a_i\right) \qquad \text{where } \mu = \mathbb{E}X_t, \tag{4}$$

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(T-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(T-1) & \gamma(T-2) & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_T \end{pmatrix} = \begin{pmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(h+T-1) \end{pmatrix}. \tag{5}$$

Let $\iota = (1, 1, \ldots, 1)'$, $\alpha_T = (a_1, \ldots, a_T)$, $\gamma_T(h) = (\gamma(h), \ldots, \gamma(h+T-1))$, and $\Gamma_T = [\gamma(i-j)]_{i,j=1,\ldots,T}$ denote the symmetric $T \times T$ covariance matrix of $(X_T, \ldots, X_1)'$, the normal equations can be written as

$$a_0 = \mu(1 - \iota'\alpha_T) \tag{6}$$

$$\Gamma_T \alpha_T = \gamma_T(h). \tag{7}$$

Dividing the second equation by $\gamma(0)$, one obtain

$$R_T \alpha_T = \rho_T(h), \tag{8}$$

where $R_T = \Gamma_T/\gamma(0)$ and $\rho(h) = (\rho(h), \ldots, \rho(h+T-1))'$. The coefficients α_T is obtained by inverting Γ_T or R_T:

$$\alpha_T = \begin{pmatrix} a_1 \\ \vdots \\ a_T \end{pmatrix} = \Gamma_T^{-1}\gamma_T(h) = R_T^{-1}\rho_T(h).$$

A sufficient condition that ensures invertibility of Γ_T and R_T is $\gamma(0) > 0$ and $\lim_{h\to\infty}\gamma(h) = 0$.

The mean squared error or *variance of the forecast error* is

$$\begin{aligned} v_T(h) &= \mathbb{E}(X_{T+h} - \mathbb{P}_T X_{T+h}))^2 \\ &= \cdots \\ &= \gamma(0) - \alpha_T'\gamma_T(h). \end{aligned}$$

3.1.1 Forecasting AR(p) Process

For AR(1) process, Eq. (5) yields (remember $\gamma(h) = \phi^h$)

$$\alpha_T = (a_1, a_2, \ldots, a_T)' = (\phi^h, 0, \ldots, 0)'.$$

We therefore get the following predictor:

$$\mathbb{P}_T X_{T+h} = \phi^h X_T.$$

For example, for $h = 1$, the forecast is ϕX_T, for $h = 2$, the forecast is $\phi(\phi X_T) = \phi^2 X_T$ etc. The variance of the forecast error is given by

$$v_T(h) = \frac{1-\phi^{2h}}{1-\phi^2}\sigma^2.$$

For $h = 1$ this is simply σ^2, and as $h \to \infty$, $v_T(h) \to \frac{1}{1-\phi^2}\sigma^2$, the unconditional variance of X_t.

For AR(p) process,it can be shown that the one-step ahead forecast is

$$\mathbb{P}_T X_{T+1} = \phi_1 X_T + \phi_2 X_{T-1} + \cdots + \phi_p X_{T+1-p}, \quad T > p.$$

The forecast for $h > 1$ can be obtained by recursively applying the forecast operator. For example, the two-step ahead forecast is

$$\begin{aligned} \mathbb{P}_T X_{T+2} &= \mathbb{P}_T(\phi_1 X_{T+1}) + \mathbb{P}_T(\phi_2 X_T) + \cdots + \mathbb{P}_T(\phi_p X_{T+2-p}) + \mathbb{P}_T(Z_{T+2}) \\ &= \phi_1(\phi_1 X_T + \phi_2 X_{T-1} + \cdots + \phi_p X_{T+1-p}) + \phi_2 X_T + \cdots + \phi_p X_{T+2-p} \\ &= (\phi_1^2 + \phi_2)X_T + (\phi_1\phi_2 + \phi_3)X_{T-1} + \cdots + (\phi_1\phi_{p-1} + \phi_p)X_{T+2-p} + (\phi_1\phi_p)X_{T+1-p}. \end{aligned}$$

3.1.2 Forecasting with MA(q) Process

Forecasting with MA(q) process is more complicated. For example, for an MA(1) process $X_t = Z_t + \theta Z_{t-1}$ with $|\theta| < 1$ and $Z_t \sim WN(0, \sigma^2)$, the one-step ahead forecast $\mathbb{P}_T X_{T+1}$ will depends on all available information $X_T, X_{T-1}, \ldots, X_1$ (namely $a_1, \ldots, a_T$ are generally non-zero). In contrast, the forecast for AR(p) model only depends on the first p observations. The coefficients of the forecast $\alpha_T = (a_1, \ldots, a_T)$ are also constant, while for the MA(q) process the forecast coefficients can change when new information arrives.

As more information becomes available, the variance of the forecast error declines monotonically. In the $h = 1$ case it will converge to σ^2. To see this, note the forecast is $\mathbb{P}_T X_{T+1} = \mathbb{P}_T Z_{T+1} + \theta \mathbb{P}_T Z_T = \theta \mathbb{P}_T Z_T$. The forecast error is Z_{T+1}. As more and more observation becomes available, it becomes more and more possible to recover the "true" value of the unobserved Z_T from the observations $X_T, X_{T-1}, \ldots, X_1$. As the process is invertible, in the limit it is possible to recover the value of Z_T exactly. The only uncertainty remaining is with respect to Z_{T+1} which has a mean of zero and a variance of σ^2.

3.2 Partial Autocorrelation Function

Recall that in our prediction problem we predict X_{T+1} as a linear combination of past data $X_T, \ldots, X_1$:

$$X_{T+1} = \mathbb{P}_T X_{T+1} + Z_{T+1} = a_1 X_T + \ldots + a_T X_1 + Z_{T+1}$$

where Z_{T+1} denote the forecast error. We can view this as a regression equation: a_T is how much X_1 contributes to the forecast of X_{T+1} after controlling for $X_T, \ldots, X_2$. We refer to a_T as the *partial autocorrelation*. In AR(p) process the information useful for forecasting $X_{T+1} (T > p)$ is incorporated in the last p observations so that $a_T = 0$. In MA process, every new observation contributes to the recovering of the Z_t's. Thus the partial autocorrelation a_T is not zero. Taking T successively to 0, 1, 2, etc, we get the partial autocorrelation function (PACF).

Definition 3.1 (First Definition)**.** The *partial autocorrelation function (PACF)* $\alpha(h)$ of a stationary process is defined as

$$\begin{aligned} \alpha(0) &= 1 \\ \alpha(h) &= a_h, \quad h = 1, 2, \ldots \end{aligned}$$

where a_h denotes the last elements of the vector $\alpha_h = \Gamma_h^{-1} \gamma_h(1) = R_h^{-1} \rho_h(1)$.

Definition 3.2 (Second Definition)**.** The *partial autocorrelation function (PACF)* $\alpha(h)$ of a stationary process is defined as

$$\begin{aligned} \alpha(0) &= 1 \\ \alpha(1) &= \text{corr}(X_2, X_1) = \rho(1) \\ \alpha(h) &= \text{corr}[X_{h+1} - \mathbb{P}(X_{h+1}|1, X_2, \ldots, X_h), X_1 - \mathbb{P}(X_1|1, X_2, \ldots, X_h)]. \end{aligned}$$

3.2.1 Autoregressive Process

The idea of the PACF can be well illustrated in the case of an AR(1) process $X_t = \phi X_{t-1} + Z_t$. X_t and X_{t-2} are correlated with each other despite the fact that there is no direct relationship between the two. The correlation is obtained "indirectly" because X_t is correlated with X_{t-1} which is itself correlated with X_{t-2}. Because both correlation are equal to ϕ, the correlation between X_t and X_{t-2} is equal to $\phi(2) = \phi^2$. The ACF therefore accounts for all correlation, including the indirect ones. The partial autocorrelation on the other hand only accounts for the direct relationships. In the case of the AR(1) process, there is only an indirect relation between X_t and $X_{t-h}$for $h \geq 2$, thus the PACF is zero.

Definition 3.1 implies that for an AR(1) process

$$\begin{aligned} \alpha_1 = \phi \quad &\Rightarrow \quad \alpha(1) = \rho(1) = \phi, \\ \alpha_2 = (\phi, 0)' \quad &\Rightarrow \quad \alpha(2) = 0, \\ \alpha_3 = (\phi, 0, 0)' \quad &\Rightarrow \quad \alpha(3) = 0. \end{aligned}$$

The PACF for an AR(1) process is therefore equal to zero for $h \geq 2$. In general, the PACF for a causal AR(p) process is equal to zero for $h > p$.

3.2.2 Moving-Average Process

Consider the case of an invertible MA process. We have

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \quad \Rightarrow \quad X_t = -\sum_{j=1}^{\infty} \pi_j X_{t-j} + Z_t.$$

X_t is therefore directly correlated with each $X_{t-h}, h = 1, 2, \ldots$. Consequently, the PACF is never exactly equal to zero, but converges exponentially to zero. This convergence can be monotonic or oscillating.

4 Estimation of Mean and ACF

4.1 Estimation of Mean

The estimation is

$$\overline{X}_T = \frac{1}{T}(X_1 + \cdots + X_T).$$

Theorem 4.1 (Convergence)**.** Suppose $\{X_t\}$ is stationary with mean μ and ACF $\gamma(h)$, then

$$\mathbb{V}\overline{X}_T = \mathbb{E}(\overline{X}_T - \mu)^2 \to 0, \qquad \text{if } \gamma(T) \to 0;$$

$$T\,\mathbb{V}\overline{X}_T = T\,\mathbb{E}(\overline{X}_T - \mu)^2 \to \sum_{h=-\infty}^{\infty} \gamma(h), \qquad \text{if } \sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$$

as $T \to \infty$.

Theorem 4.2 (Asymptotic Distribution of Sample Mean)**.** For any stationary process $\{X_t\}$ given by

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad Z_t \sim IID(0, \sigma^2)$$

such that $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $\sum_{j=-\infty}^{\infty} \psi_j \neq 0$, the average $\overline{X}_T$ is asymptotically normal:

$$\begin{aligned}\sqrt{T}(\overline{X}_T - \mu) &\rightsquigarrow N\left(0, \sum_{h=-\infty}^{\infty} \gamma(h)\right) \\ &= N\left(0, \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j\right)^2\right) = N(0, \sigma^2 \Psi(1)^2).\end{aligned}$$

4.2 Estimation of the Autocovariance and Autocorrelation Function

The estimation is

$$\hat{\gamma}(h) = \frac{1}{T} \sum_{t=1}^{T-h} (X_t - \overline{X}_T)(X_{t+h} - \overline{X}_T) \tag{9}$$

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}. \tag{10}$$

Theorem 4.3 (Asymptotic Distribution of Autocorrelations)**.** For any stationary process $\{X_t\}$ given by

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad Z_t \sim IID(0, \sigma^2)$$

such that $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $\sum_{j=-\infty}^{\infty} j\,|\psi_j|^2 < \infty$,

$$\begin{pmatrix} \hat{\rho}(1) \\ \vdots \\ \hat{\rho}(h) \end{pmatrix} \rightsquigarrow N\left(\begin{pmatrix} \rho(1) \\ \vdots \\ \rho(h) \end{pmatrix}, \frac{W}{T} \right)$$

where $W = (w_{ij})_{i,j=1,\dots,h}$ is given by

$$w_{ij} = \sum_{k=1}^{\infty} [\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)][\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)].$$

When $\{X_t\}$ is a white noise process, $\rho(h) = 0$ for $|h| > 0$, and Theorem 4.3 implies that W is the identity matrix I_h. The asymptotic distribution of $\sqrt{T}\hat{\rho}(h)$ converges to the standard normal distribution $N(0,1)$. This implies that for large T we can approximate the distribution of $\hat{\rho}(h)$ by a normal distribution with mean zero and variance $1/T$. This allows the construction of a 95% confidence interval assuming that the true process is a white noise. This confidence interval is therefore given by $\pm 1.96 T^{-1/2}$. For $T = 100$ this is ± 0.196.

Instead of examining each correlation coefficient separately, we can test the joint hypothesis that all correlation coefficients up to order N are simultaneously equal to zero, i.e. $\rho(1) = \rho(2) = \cdots = \rho(N) = 0$. As each $\sqrt{T}\hat{\rho}(h)$ has an asymptotic standard normal distribution, the sum of the squared autocorrelation coefficients is χ^2 distributed with N degrees of freedom. This test statistic is called *Box-Pierce* statistic:

$$Q = T \sum_{h=1}^{N} \hat{\rho}^2(h) \quad \sim \quad \chi^2_N.$$

A refinement of the test is *Ljung-Box Statistic*:

$$Q' = T(T+2) \sum_{h=1}^{N} \frac{\hat{\rho}^2(h)}{T-h} \quad \sim \quad \chi^2_N.$$

This statistic gives more weights to smaller h and less weights to higher h, as for higher h the number of observations used to calculate $\hat{\rho}(h)$ is small. We reject the null hypothesis that all correlation coefficients are jointly equal to zero if the statistic is very large (p-value very small).

When $\{X_t\}$ is a MA(q) process or an AR(p) process, confidence intervals can be constructed in a similar fashion according to Theorem 4.3.

4.3 Estimation of the Partial Autocorrelation Function

According to Definition 3.1 α_h and consequently a_h can be estimated by $\widehat{\alpha}_h = \widehat{\Gamma}_h^{-1}\hat{\gamma}_h(1) = \widehat{R}_h^{-1}\hat{\rho}_h(1)$. Since $\hat{\rho}(h)$ is asymptotically normally distributed, we have that $\widehat{\alpha}_h$ is also asymptotically normal. In particular for AR(p) process

$$\sqrt{T}\widehat{\alpha}(h) \rightsquigarrow N(0,1) \quad \text{for } T \to \infty \text{ and } h > p.$$

The result allows us to construct confidence intervals for the estimated partial autocorrelation coefficients.

4.4 Estimation of the Long-Run Variance

The long-run variance J is

$$J = \sum_{h=\infty}^{\infty} \gamma(h) = \gamma(0) + 2\sum_{h=1}^{\infty} \gamma(h) = \gamma(0)\left(1 + 2\sum_{h=1}^{\infty} \rho(h)\right).$$

A first naive estimate of J is $\widehat{J}_T$ defined as

$$\widehat{J}_T = \sum_{h=-T+1}^{T-1} \hat{\gamma}(h).$$

However, estimates of higher order autocovariances are based on smaller samples, so their estimates become less reliable. We can see only a certain number ℓ_T of autocovariances and/or use a weighted sum:

$$\widehat{J}_T = \widehat{J}_T(\ell_T) = \frac{T}{T-r} \sum_{h=-T+1}^{T-1} k\left(\frac{h}{\ell_T}\right) \hat{\gamma}(h)$$

where k is a kernel function. The kernel functions are required to have the following properties:

1. $k : \mathbb{R} \to [-1, 1]$ is continuous except for a finite number of points. In particular k should be continuous at $x = 0$.
2. $\|k\|_2 = \int_{\mathbb{R}} k^2(x)dx < \infty$.
3. $k(0) = 1$.
4. k is symmetric, i.e. $k(x) = k(-x)$ for all $x \in \mathbb{R}$.

See Fig. 1 for common kernel functions[1].

[1] From the textbook.

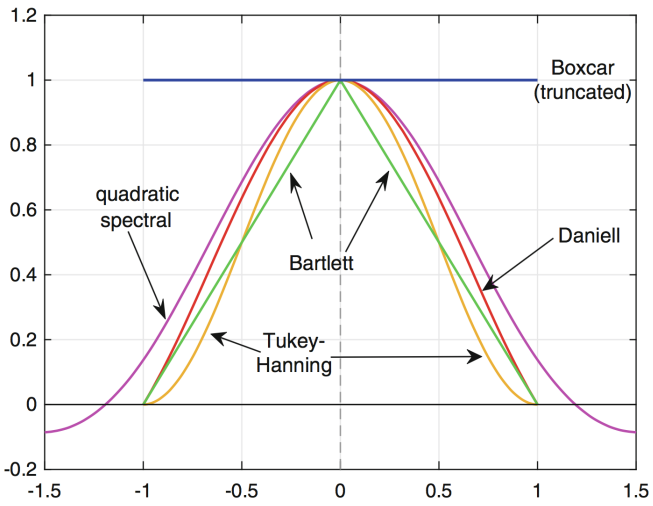

Figure 1: common kernel functions

5 Estimation of ARMA Models

In this section, we talk about the issue of estimating the coefficients of ARMA models from data. We show that for AR(p) models we can use the Yule-Walker estimation or the OLS estimation, while for the general ARMA models we have to resort to Maximum Likelihood estimation.

5.1 The Yule-Walker Estimator

For AR models, we can use the estimated autocovariance functions to solve for the coefficients: for

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

multiply it by $X_t, X_{t-1}, \ldots, X_{t-p}$ and taking expectations respectively lead to the following system of linear equations for $\Phi = (\phi_1, \ldots, \phi_p)'$ and σ^2:

$$\begin{aligned} \gamma(0) - \phi_1\gamma(1) - \cdots - \phi_p\gamma(p) &= \sigma^2 \\ \gamma(1) - \phi_1\gamma(0) - \cdots - \phi_p\gamma(p-1) &= 0 \\ &\cdots \\ \gamma(p) - \phi_1\gamma(p-1) - \cdots - \phi_p\gamma(0) &= 0. \end{aligned}$$

This equation system is known as the *Yule-Walker equations*. If can be written in matrix algebra as

$$\gamma(0) - \Phi'\gamma_p(1) = \sigma^2,$$

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix},$$

or

$$\begin{aligned} \gamma(0) - \Phi'\gamma_p(1) &= \sigma^2, \\ \Gamma_p\Phi &= \gamma_p(1). \end{aligned}$$

The estimator is

$$\widehat{\Phi} = \widehat{\Gamma}_p^{-1}\hat{\gamma}_p(1) = \widehat{R}_p^{-1}\hat{\rho}_h(1),$$

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \widehat{\Phi}'\hat{\gamma}_p(1).$$

Theorem 5.1 (Asymptotic Normality of Yule-Walker Estimator)**.** Let $\{X_t\}$ be an AR(p) process which is causal with respect to $\{Z_t\} \sim IID(0, \sigma^2)$. The Yule-Walker estimator is consistent and $\widehat{\Phi}$ is asymptotically normal:

$$\sqrt{T}\left(\widehat{\Phi} - \Phi\right) \rightsquigarrow N\left(0, \sigma^2\Gamma_p^{-1}\right).$$

In addition we have

$$\hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

For an AR(1) process, namely $p = 1$, we have $\Gamma_1 = \gamma(0)$, $\gamma_1(1) = \gamma(1)$, and the equation simplifies to $\gamma(0)\phi = \gamma(1)$. The estimator for ϕ is $\hat{\phi} = \hat{\gamma}(1)/\hat{\gamma}(0) = \hat{\rho}(1)$. The asymptotic distribution is

$$\sqrt{T}\left(\hat{\phi} - \phi\right) \rightsquigarrow N\left(0, \frac{\sigma^2}{\gamma(0)}\right) = N(0, 1 - \phi^2).$$

(Recall that $\gamma(0)$ is $\sigma^2/(1-\phi^2)$)

We mention that the Yule-Walker estimator is not suitable for the MA process, since it is no longer consistent and the system of equations is no longer linear.

5.2 OLS Estimation of AR(p) Model

We can view the AR model as a regression model:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t, \qquad Z_t \sim WN(0, \sigma^2).$$

Given the data $\{X_1, \ldots, X_T\}$, we can write out the equation for each $X_{p+1}, \ldots, X_T$:

$$\begin{pmatrix} X_{p+1} \\ X_{p+2} \\ \vdots \\ X_T \end{pmatrix} = \begin{pmatrix} X_p & X_{p-1} & \cdots & X_1 \\ X_{p+1} & X_p & \cdots & X_2 \\ \vdots & \vdots & \ddots & \vdots \\ X_{T-1} & X_{T-2} & \cdots & X_{T-p} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} + \begin{pmatrix} Z_{p+1} \\ Z_{p+2} \\ \vdots \\ Z_T \end{pmatrix}$$

or

$$Y = X\Phi + Z.$$

The solution is given by

$$\widehat{\Phi} = (X'X)^{-1}(X'Y).$$

The standard orthogonality assumption between regressors and error is violated. Thus the OLS estimator is not unbiased in finite samples, although $\sqrt{T}\left(\widehat{\Phi} - \Phi\right)$ is asymptotically normal:

Theorem 5.2. Under the same assumptions as in Theorem 5.1, the OLS estimator is asymptotically distributed as

$$\sqrt{T}\left(\widehat{\Phi} - \Phi\right) \rightsquigarrow N(0, \sigma^2\Gamma_p^{-1}),$$

and

$$s_T^2 \xrightarrow{p} \sigma^2$$

where $s_T^2 = \widehat{Z}'\widehat{Z}/T$ and $\widehat{Z}_t$ is the OLS residual.

In practice, $\sigma^2\Gamma_p^{-1}$ is approximated by $s_T^2(X'X/T)^{-1}$. Thus for large T, $\widehat{\Phi}$ can be viewed as being normally distributed as $N(\Phi, s_T^2(X'X)^{-1})$. This result allows the application of the usual t and F tests.

5.3 Maximum Likelihood Estimation of ARMA(p, q) Model

We can not use the OLS estimation directly in case of ARMA(p, q) models, since the MA part $Z_{t-1}, \ldots, Z_{t-q}$ are not directly observable. Instead, we resort to maximum likelihood estimate. To do that, it is necessary to assume a likelihood function of the data. We assume that the data $(X_1, \ldots, X_T)'$ is distributed as a multivariate normal with mean zero and variance Γ_T. The parameters are $\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ and σ^2. The Gaussian likelihood function is given by

$$L_T(\beta, \sigma^2 | x_T) = (2\pi)^{-T/2} (\det \Gamma_T)^{-1/2} \exp\left(-\frac{1}{2} x_T' \Gamma_T^{-1} x_T\right)$$

where $x_T = (x_1, \ldots, x_T)$ is the observed data. We can then choose β and σ^2 to maximize the likelihood function.

Theorem 5.3. If $\{X_t\}$ is an ARMA(p, q) process with true parameters β and $Z_t \sim IID(0, \sigma^2)$ then the maximum-likelihood estimator has asymptotically normal distribution:

$$\sqrt{T}\left(\hat{\beta}_{ML} - \beta\right) \rightsquigarrow N(0, V(\beta)).$$

The asymptotic covariance matrix $V(\beta)$ is given by

$$V(\beta) = \begin{pmatrix} \mathbb{E} U_t U_t' & \mathbb{E} U_t V_t' \\ \mathbb{E} U_t V_t' & \mathbb{E} V_t V_t' \end{pmatrix}^{-1}.$$

U_t is $(u_t, \ldots, u_{t-p+1})$ and V_t is $(v_t, \ldots, v_{t-q+1})$ where $\{u_t\}$ and $\{v_t\}$ denote autoregressive processes defined as $\Phi(L) u_t = w_t$ and $\Theta(L) v_t = w_t$ with $w_t \sim WN(0, 1)$.

5.4 Estimation of the Orders p and q

Three criteria: Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the Hannan-Quinn information criterion (HQ criterion):

$$AIC(p, q) = \ln \hat{\sigma}_{p,q}^2 + (p + q)\frac{2}{T}$$

$$BIC(p, q) = \ln \hat{\sigma}_{p,q}^2 + (p + q)\frac{\ln T}{T}$$

$$HQC(p, q) = \ln \hat{\sigma}_{p,q}^2 + (p + q)\frac{2 \ln(\ln(T))}{T}$$

where $\hat{\sigma}_{p,q}^2$ is the variance of the residuals from an estimate of ARMA(p, q). One chooses p and q so as to minimize the criteria. We have $AIC < HQC < BIC$ for $T \geq 16$, so AIC delivers the largest models, while BIC delivers the smallest models.

5.5 Modeling a Stochastic Process

- **Step 1: Transformations to achieve stationarity:**
  - Take differences: $Y_t = (1 - \mathrm{L})^d X_t$. If $\{Y_t\}$ is stationary then $\{X_t\}$ is said to be integrated of order d, and we write $X_t \sim I(d)$. If $\{Y_t\}$ is ARMA(p, q) then we call $\{X_t\}$ ARIMA(p, d, q) process.
  - Seasonal difference: in the case of quarterly observations $Y_t = (1 - \mathrm{L}^4)X_t$. Since $1 - \mathrm{L}^4 = (1 - \mathrm{L})(1 + \mathrm{L} + \mathrm{L}^2 + \mathrm{L}^3)$, this transformation can also account for the trend.
  - Filters.
- **Step 2: Finding the Orders p and q:** We can either analyze ACF and PACF, or use the information criteria.
- **Step 3: Checking Plausibility:** After having identified a particular model, we can check:

  (i) Are the residuals white noise? This can be checked by investigating the ACF of the residuals or by applying the Ljung-Box test. If they are not this means that the model failed to capture all the dynamics in the data.

  (ii) Are the parameters plausible?

  (iii) Are the parameters constant over time? Are there structural breaks?

  (iv) Does the model deliver sensible forecasts? It is useful to investigate the out-of-sample forecast performance.

6 Integrated Processes

6.1 Definition and Properties

- *Trend-stationary* process:

$$X_t = \underbrace{\alpha + \delta t}_{\text{linear trend}} + \Psi(\mathrm{L}) Z_t.$$

- *Difference-stationary* process:

Definition 6.1. The stochastic process $\{X_t\}$ is called *integrated of order one* or *difference-stationary*, denoted as $X_t \sim I(1)$, if $\Delta X_t = X_t - X_{t-1}$ can be represented as

$$\Delta X_t = (1 - \mathrm{L}) X_t = \delta + \Psi(\mathrm{L}) Z_t, \qquad \Psi(1) \neq 0$$

with $\{Z_t\} \sim WN(0, \sigma^2)$ and $\sum_{j=0}^{\infty} j\,|\psi_j| < \infty$.

6.1.1 Long-Run Forecast

For the trend-stationary process, the least-squares forecast given the infinite past is

$$\widetilde{\mathbb{P}}_t X_{t+h} = \alpha + \delta(t+h) + \psi_h Z_t + \psi_{h+1} Z_{t-1} + \cdots$$

We have

$$\lim_{h\to\infty} \mathbb{E}\left(\widetilde{\mathbb{P}}_t X_{t+h} - \alpha - \delta(t+h)\right)^2 = \sigma^2 \lim_{h\to\infty} \sum_{j=0}^{\infty} \psi_{h+j}^2 = 0.$$

Thus the long-run forecast is the linear trend. Even if X_t deviates temporarily from the trend line, it is assumed to return to it.

The forecast for the difference-stationary process is

$$\widetilde{\mathbb{P}}_t \Delta X_{t+h} = \delta + \psi_h Z_t + \psi_{h+1} Z_{t-1} + \cdots$$

The level of X_{t+h} is

$$X_{t+h} = (X_{t+h} - X_{t+h-1}) + (X_{t+h-1} - X_{t+h-2}) + \cdots + (X_{t+1} - X_t) + X_t$$

so that

$$\begin{aligned}
\widetilde{\mathbb{P}}_t X_{t+h} &= \widetilde{\mathbb{P}}_t \Delta X_{t+h} + \widetilde{\mathbb{P}}_t \Delta X_{t+h-1} + \cdots + \widetilde{\mathbb{P}}_t \Delta X_{t+1} + X_t \\
&= \cdots \\
&= X_t + \delta h + (\psi_h + \cdots + \psi_1) Z_t + (\psi_{h+1} + \cdots + \psi_1) Z_{t-1} + \cdots
\end{aligned}$$

This shows that the intercept of the long-run forecast X_t is no longer a fixed number, but is stochastic. For a random walk with drift $\{X_t\}$, the best forecast for X_{t+h} is $\mathbb{P}_t X_{t+h} = \delta h + X_t$.

6.1.2 Variance of Forecast Error

For trend-stationary process the forecast error is

$$X_{t+h} - \widetilde{\mathbb{P}}_t X_{t+h} = Z_{t+h} + \psi_1 Z_{t+h-1} + \cdots + \psi_{h-1} Z_{t+1}.$$

The variance is

$$\mathbb{E}\left(X_{t+h} - \widetilde{\mathbb{P}}_t X_{t+h}\right)^2 = (1 + \psi_1^2 + \cdots + \psi_{h-1}^2)\sigma^2.$$

For $h \to \infty$ this converges to $\sigma^2 \sum_{j=0}^{\infty} \psi_j^2 < \infty$. This is nothing more than the unconditional variance of X_t.

It can be shown that for the integrated process the variance is

$$\mathbb{E}\left(X_{t+h} - \widetilde{\mathbb{P}}_t X_{t+h}\right)^2 = \left[1 + (1+\psi_1)^2 + \cdots + (1 + \psi_1 + \cdots + \psi_{h-1})^2\right]\sigma^2$$

which goes to infinity as $h \to \infty$.

6.1.3 Impulse Response Function

For the trend-stationary process

$$\frac{\partial \widetilde{\mathbb{P}}_t X_{t+h}}{\partial Z_t} = \psi_h \to 0 \qquad \text{as } h \to \infty.$$

The effect of a shock thus declines with time and dies out. Shocks have therefore only transitory or temporary effects. In the case of an ARMA process the effect declines exponentially. In the case of integrated process

$$\frac{\partial \widetilde{\mathbb{P}}_t X_{t+h}}{\partial Z_t} = 1 + \psi_1 + \psi_2 + \cdots + \psi_h.$$

As $h \to \infty$, this converges to $\sum_{j=0}^{\infty} \psi_j = \Psi(1) \neq 0$. Thus a shock will have a long-run or permanent effect. This long-run effect is called *persistence.* If $\{\Delta X_t\}$ is an ARMA process then the persistence is

$$\Psi(1) = \frac{\Theta(1)}{\Phi(1)}.$$

6.1.4 The Beveridge-Nelson Decomposition

Theorem 6.2 (Beveridge-Nelson Decomposition)**.** Every integrated process $\{X_t\}$ can be decomposed as

$$X_t = \underbrace{X_0 + \delta t}_{\text{linear trend}} + \underbrace{\Psi(1) \sum_{j=1}^{t} Z_j}_{\text{random walk}} + \underbrace{\widetilde{\Psi}(\mathrm{L}) Z_0 - \widetilde{\Psi}(\mathrm{L}) Z_t}_{\text{stationary component}}.$$

Proof. First observe that

$$\begin{aligned}\Psi(\mathrm{L}) - \Psi(1) &= (1 + \psi_1\mathrm{L} + \psi_2\mathrm{L}^2 + \cdots) - (1 + \psi_1 + \psi_2 + \cdots) \\ &= \psi_1(\mathrm{L} - 1) + \psi_2(\mathrm{L}^2 - 1) + \cdots \\ &= (\mathrm{L} - 1)[\psi_1 + \psi_2(\mathrm{L} + 1) + \psi_3(\mathrm{L}^2 + \mathrm{L} + 1) + \cdots] \\ &= (\mathrm{L} - 1)\left[(\psi_1 + \psi_2 + \cdots) + (\psi_2 + \psi_3 + \cdots)\mathrm{L} + (\psi_3 + \psi_4 + \cdots)\mathrm{L}^2 + \cdots\right].\end{aligned}$$

Thus we have $\Psi(\mathrm{L}) = \Psi(1) + (\mathrm{L} - 1)\widetilde{\Psi}(\mathrm{L})$ where $\widetilde{\Psi}(\mathrm{L}) = \sum_{j=0}^{\infty} \widetilde{\psi}_j L^j$ with $\widetilde{\psi}_j = \sum_{i=j+1}^{\infty} \psi_i$. We can then express X_t as

$$\begin{aligned}X_t &= X_0 + \sum_{j=1}^{t} \Delta X_j \\ &= X_0 + \sum_{j=1}^{t} \left\{\delta + \left[\Psi(1) + (\mathrm{L} - 1)\widetilde{\Psi}(\mathrm{L})\right] Z_j\right\} \\ &= X_0 + \delta t + \Psi(1)\sum_{j=1}^{t} Z_j + \sum_{j=1}^{t} (\mathrm{L} - 1)\widetilde{\Psi}(\mathrm{L}) Z_j \\ &= \underbrace{X_0 + \delta t}_{\text{linear trend}} + \underbrace{\Psi(1)\sum_{j=1}^{t} Z_j}_{\text{random walk}} + \underbrace{\widetilde{\Psi}(\mathrm{L})Z_0 - \widetilde{\Psi}(\mathrm{L})Z_t}_{\text{stationary component}}.\end{aligned}$$

To prove the last component is stationary, we need to show $\sum_{j=0}^{\infty} |\widetilde{\psi}_j| < \infty$. Indeed

$$\sum_{j=0}^{\infty} |\widetilde{\psi}_j| < \infty = \sum_{j=0}^{\infty} \left| \sum_{i=j+1}^{\infty} \psi_i \right| \leq \sum_{j=0}^{\infty} \sum_{i=j+1}^{\infty} |\psi_i| = \sum_{j=1}^{\infty} j\,|\psi_j| \leq \infty$$

where the last inequality follows from Definition 6.1. □

Note that the coefficient of the random walk component, $\Psi(1)$, is the persistence. In macroeconomics aggregate supply shocks are ascribed to have a long-run effect as they affect productivity. In contrast monetary or demand shocks are viewed to have temporary effects only. Thus the persistence $\Psi(1)$ can be interpreted as a measure for the importance of supply shocks.

6.2 Unit-Root Tests

For an AR(1) model $X_t = \phi X_{t-1} + Z_t$, if $|\phi| < 1$, then

$$\sqrt{T}\left(\hat{\phi}_T - \phi\right) \rightsquigarrow N(0, 1 - \phi^2).$$

If $\phi = 1$, then

$$T\left(\hat{\phi}_T - \phi\right) \rightsquigarrow \nu$$

where ν is the Dickey-Fuller distribution. To determine whether $\phi = 1$, we use the Dickey-Fuller regression

$$X_t = \text{deterministic variables} + \phi X_{t-1} + Z_t.$$

An alternatively and numerically equivalent regression is

$$\Delta X_t = \text{deterministic variables} + \beta X_{t-1} + Z_t$$

where $\beta = \phi - 1$. The tests are

$$H_0 : \phi = 1 \quad \text{v.s.} \quad -1 < \phi < 1$$

or

$$H_0 : \beta = 0 \quad \text{v.s.} \quad -2 < \beta < 0.$$

The test statistic can be $T\left(\hat{\phi}_T - 1\right)$ but a much more common choice is

$$t = (\hat{\phi}_T - 1)/\hat{\sigma}_{\hat{\phi}}.$$

This statistic is also not asymptotically normally distributed, and its distribution can be found in time series textbooks. If we control for lagged differenced in the regression, as in

$$X_t = \text{deterministic variables} + \phi X_{t-1} + \gamma_1 \Delta X_{t-1} + \cdots + \gamma_{p-1} \Delta X_{t-p+1} + Z_t$$

then the test is the called the augmented Dickey-Fuller test (ADF-test). This autoregressive correction does not change the asymptotic distribution of the test statistic, so that the same table can be used. The γ_i's are asymptotically normal.

Part II

Multivariate Time Series Analysis

7 Definition and Stationarity

We view X_t as $(X_{1,t}, \ldots, X_{n,t})'$. The first two moments: $\mathbb{E}X_t = \mu_t$ and

$$\Gamma(t,s) = \begin{pmatrix} \gamma_{11}(t,s) & \cdots & \gamma_{1n}(t,s) \\ \vdots & \ddots & \vdots \\ \gamma_{n1}(t,s) & \cdots & \gamma_{nn}(t,s) \end{pmatrix}$$

where $\gamma_{ij}(t,s) = \mathbb{E}(X_{i,t} - \mu_{i,t})(X_{j,s} - \mu_{j,s}), \quad i,j = 1,\ldots,n$. Stationarity as the invariance in the first two moments to time shifts:

Definition 7.1. A multivariate stochastic process $\{X_t\}$ is *stationary* if

1. $\mathbb{E}X_t$ is constant;
2. $\mathbb{E}X_t'X_t < \infty$;
3. $\Gamma(t,s) = \Gamma(t+r, s+r)$.

We define the correlation matrix $R(h) = (\rho_{ij}(h))$ as

$$\rho_{ij}(h) = \frac{\gamma_{ij}(h)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}}.$$

The correlation matrix can be written as

$$R(h) = V^{-1/2}\Gamma(h)V^{-1/2}$$

where V is a diagonal matrix with diagonal elements $\gamma_{ii}(0)$. Clearly the diagonal of $R(h)$ is 1. Note that in general $\rho_{ij}(h) \neq \rho_{ji}(h)$ for $h \neq 0$, and it is possible that $\rho_{ij}(h) > \rho_{ij}(0)$. Taking the following as an example:

$$\begin{aligned} X_{1t} &= Z_t \\ X_{2t} &= Z_t + 0.75Z_{t-2} \end{aligned}$$

with $Z_t \sim WN(0,1)$. We have $\mathbb{E}X_t = 0$ and $\Gamma(0) = \begin{pmatrix} 1 & 1 \\ 1 & 1.5625 \end{pmatrix}$, $\Gamma(1) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, and $\Gamma(2) = \begin{pmatrix} 0 & 0 \\ 0.75 & 0.75 \end{pmatrix}$. The correlation function is $R(0) = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$, $R(1) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, and $R(2) = \begin{pmatrix} 0 & 0 \\ 0.60 & 0.48 \end{pmatrix}$.

Definition 7.2. A stochastic process $\{Z_t\}$ is called *multivariate white noise process* with mean zero and covariance matrix $\Sigma > 0$, denoted by $Z_t \sim WN(0, \Sigma)$, if it is stationary, $\mathbb{E}Z_t = 0$, and

$$\Gamma(h) = \begin{cases} \Sigma & h = 0; \\ 0 & h \neq 0. \end{cases}$$

If $\{Z_t\}$ is i.i.d. then we write $Z_t \sim IID(0, \Sigma)$.

Note that for the process $Z_t = (u_t, u_{t-1})$, where $u_t \sim WN(0, \sigma_u^2)$, we have $\Gamma(1) = \begin{pmatrix} 0 & 0 \\ \sigma_u^2 & 0 \end{pmatrix} \neq 0$, so that it is not white noise according to our definition, even if each component is a white noise process.

Taking moving averages of a white noise process it is possible to generate new stationary processes. This leads to the definition of a linear process.

Definition 7.3. A stochastic process $\{X_t\}$ is called *linear* if

$$X_t = \sum_{j=-\infty}^{\infty} \Psi_j Z_{t-j}$$

where $Z_t \sim IID(0, \Sigma)$ and $\sum_{j=-\infty}^{\infty} \|\Psi_j\| < \infty$. If $\Psi_j = 0$ for all $j < 0$ then it is called an VMA(∞) process.

8 Estimation of Mean and Covariance Function

The estimation of mean is $\hat{\mu} = \overline{X}_T = \frac{1}{T}(X_1 + \cdots + X_T) = (\overline{X}_1, \ldots, \overline{X}_n)'$. The estimation is unbiased and consistent:

Theorem 8.1. As $T \to \infty$:

$$\mathbb{E}(\overline{X}_T - \mu)'(\overline{X}_T - \mu) \to 0 \quad \text{if } \gamma_{ii}(T) \to 0 \text{ for all } 1 \leq i \leq n$$

and

$$T\,\mathbb{E}(\overline{X}_T - \mu)'(\overline{X}_T - \mu) \to \sum_{i=1}^{n} \sum_{h=-\infty}^{\infty} \gamma_{ii}(h)$$

if $\sum_{h=-\infty}^{\infty} \gamma_{ii}(h) < \infty$ for all $1 \leq i \leq n$.

The estimator is also asymptotically normally distributed:

Theorem 8.2. For any stationary process $\{X_t\}$

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \Psi_j Z_{t-j}$$

with $Z_t \sim IID(0, \Sigma)$ and $\sum_{j=-\infty}^{\infty} \|\Psi_j\| < \infty$, the average $\overline{X}_T$ is asymptotically normal:

$$\begin{aligned}
\sqrt{T}\left(\overline{X}_T - \mu\right) &\rightsquigarrow N\left(0, \sum_{h=-\infty}^{\infty} \Gamma(h)\right) \\
&= N\left(0, \left(\sum_{j=-\infty}^{\infty} \Psi_j\right) \Sigma \left(\sum_{j=-\infty}^{\infty} \Psi_j'\right)\right) \\
&= N(0, \Psi(1)\Sigma\Psi(1)').
\end{aligned}$$

For the estimation of the covariance matrix

$$\widehat{\Gamma}(h) = \begin{cases} \frac{1}{T}\sum_{t=1}^{T-h}(X_{t+h} - \overline{X}_T)(X_t - \overline{X}_T)', & 0 \leq h \leq T-1 \\ \widehat{\Gamma}'(-h), & -T+1 \leq h < 0. \end{cases}$$

The estimator of the correlation function is $\widehat{R}(h) = \widehat{V}^{-1/2}\widehat{\Gamma}(h)\widehat{V}^{-1/2}$ where $\widehat{V}^{1/2} = \text{diag}(\sqrt{\hat{\gamma}_{11}(0)}, \ldots, \sqrt{\hat{\gamma}_{nn}(0)}$
The estimator converges in probability to the true covariance matrix $\Gamma(h)$, and $\sqrt{T}\left(\widehat{\Gamma}(h) - \Gamma(h)\right)$
is asymptotically normally distributed.

Theorem 8.3. Let $\{X_t\}$ be a stationary process with

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \Psi_j Z_{t-j}$$

with $Z_t \sim IID(0, \Sigma)$, $\sum_{j=-\infty}^{\infty} \|\Psi_j\| < \infty$, and $\sum_{j=-\infty}^{\infty} \Psi_j \neq 0$. Then for each h, $\widehat{\Gamma}(h)$ converges in probability to $\Gamma(h)$:

$$\widehat{\Gamma}(h) \xrightarrow{p} \Gamma(h).$$

As in the univariate case, we can define the long-run covariance matrix J as

$$J = \sum_{h=-\infty}^{\infty} \Gamma(h).$$

We can again consider the estimator

$$\widehat{J}(T) = \sum_{h=-T+1}^{T-1} k\left(\frac{h}{\ell_T}\right) \widehat{\Gamma}(h).$$

9 VARMA Process

Definition 9.1. A multivariate stochastic process $\{X_t\}$ is a vector autoregressive moving-average process of order (p, q), denoted as VARMA(p, q), if it is stationary and

$$X_t - \Phi_1 X_{t-1} - \cdots - \Phi_p X_{t-p} = Z_t + \Theta_1 Z_{t-1} + \cdots + \Theta_q Z_{t-q}$$

where $\Phi_p \neq 0$, $\Theta_q \neq 0$, and $Z_t \sim WN(0, \Sigma)$. $\{X_t\}$ is called a VARMA(p, q) process with mean μ if $\{X_t - \mu\}$ is a VARMA(p, q) process.

We write

$$\Phi(\mathrm{L}) Z_t = \Theta(\mathrm{L}) Z_t$$

where $\Phi(\mathrm{L}) = I_n - \Phi_1 \mathrm{L} - \cdots - \Phi_p \mathrm{L}^p$ and $\Theta(\mathrm{L}) = I_n + \Theta_1 \mathrm{L} + \cdots + \Theta_q \mathrm{L}^q$. Note that $\Phi(\mathrm{L})$ is an $n \times n$ matrix whose elements are lag polynomials. Similarly $\Theta(\mathrm{L})$ is also an $n \times n$ matrix whose elements are lag polynomials.

9.1 The VAR(1) process

We now have a closer look at the VAR(1) process, i.e.

$$X_t = \Phi X_{t-1} + Z_t \qquad \text{with } Z_t \sim WN(0, \Sigma).$$

We assume all eigenvalues of the matrix Φ are absolutely strictly smaller than one. As the eigenvalues correspond to the inverses of the roots of the characteristic polynomial $\det(I_n - \Phi z)$, this assumption implies that all roots lie outside the unit circle, i.e. $\det(I_n - \Phi z) \neq 0$ for all $z \in \mathbb{C}$ with $|z| \leq 1$. It can be shown that the solution to the difference equation is

$$X_t = Z_t + \Phi_1 Z_{t-1} + \Phi^2 Z_{t-2} + \cdots = \sum_{j=0}^{\infty} \Phi^j Z_{t-j}.$$

9.2 Causal Representation

Note that the causal representation is with respect to a general VARMA(p, q) process that can include MA terms.

Definition 9.2. A VARMA(p, q) process $\{X_t\}$ with $\Phi(\mathrm{L}) X_t = \Theta(\mathrm{L}) Z_t$ is called *causal* with respect to $\{Z_t\}$ if and only if there exists a sequence of absolutely summable matrices $\{\Psi_j\}$. i.e. $\sum_{j=0}^{\infty} \|\Psi_j\| < \infty$ such that

$$X_t = \sum_{j=0}^{\infty} \Psi_j Z_{t-j}.$$

Theorem 9.3. Let $\{X_t\}$ ne a VARMA(p,q) process with $\Phi(\mathrm{L})X_t = \Theta(\mathrm{L})Z_t$ and assume that

$$\det \Phi(z) \neq 0 \qquad \text{for all } z \in \mathcal{C} \text{ with } |z| \leq 1.$$

Then the stochastic difference equation has exactly one stationary solution with causal representation

$$X_t = \sum_{j=0}^{\infty} \Psi_j Z_{t-j},$$

where $\{\Psi_j\}$ is determined by the identity

$$\Phi(z)\Psi(z) = \Theta(z).$$

Below, however, we focus on VAR processes.

9.3 Computation of the Covariance Function

Let's first consider a VAR(1) process

$$X_t = \Phi X_{t-1} + Z_t, \qquad Z_t \sim WN(0, \Sigma).$$

We multiply the above equation by X_t' and then X_{t-h}' to obtain

$$\mathbb{E}(X_t X_t') = \Gamma(0) = \Phi\mathbb{E}(X_{t-1}X_t') + \mathbb{E}(Z_t X_t') = \Phi\Gamma(-1) + \Sigma,$$

$$\mathbb{E}(X_t X_{t-h}') = \Gamma(h) = \Phi\mathbb{E}(X_{t-1}X_{t-h}') + \mathbb{E}(Z_t X_{t-h}') = \Phi\Gamma(h-1).$$

Knowing $\Gamma(0)$ and Φ, the $\Gamma(h)$ can be computed recursively as

$$\Gamma(h) = \Phi^h \Gamma(0).$$

We can solve for $\Gamma(0)$ from the first equation as follows: note that $\Gamma(-1) = \Gamma(1)' = (\Phi\Gamma(0))' = \Gamma(0)\Phi'$, and so

$$\Gamma(0) = \Phi\Gamma(0)\Phi' + \Sigma.$$

We have

$$\mathrm{vec}\Gamma(0) = \mathrm{vec}(\Phi\Gamma(0)\Phi') + \mathrm{vec}\Sigma = (\Phi \otimes \Phi)\mathrm{vec}\Gamma(0) + \mathrm{vec}\Sigma,$$

and so

$$\mathrm{vec}\Gamma(0) = (I_{n^2} - \Phi \otimes \Phi)^{-1}\mathrm{vec}\Sigma.$$

The covariance function of a causal VAR(p) process can be calculated by first transform the process into the companion form as a VAR(1) process and then applies the procedure above.

10 Estimation of VAR Models

10.1 OLS Estimation

The estimation is complicated so we just state the main results here. Notation: we suppose we have $T+p$ observations with $t = T, T-1, \ldots, 0, \ldots, -p+1$. And we write $Y = (X_1, \ldots, X_T)$, $\Phi = (\Phi_1, \ldots, \Phi_p)$, $Z = (Z_1, \ldots, Z_T)$ and

$$X = \begin{pmatrix} X_{1,0} & \cdots & X_{n,0} & \cdots & X_{1,-p+1} & \cdots & X_{n,-p+1} \\ X_{1,1} & \cdots & X_{n,1} & \cdots & X_{1,-p+2} & \cdots & X_{n,-p+2} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{1,T-1} & \cdots & X_{n,T-1} & \cdots & X_{1,T-p} & \cdots & X_{n,T-p} \end{pmatrix},$$

namely the first row runs from 0 to $-p+1$, the second row runs from 1 to $-p+2$, and so on, till the last row that runs from $T-1$ to $T-p$. The OLS estimator is

$$(\mathrm{vec}\widehat{\Phi})_{OLS} = (((X'X)^{-1}X') \otimes I_n)\mathrm{vec}Y$$

or

$$\widehat{\Phi} = YX(X'X)^{-1}.$$

Theorem 10.1 (Asymptotic Distribution of OLS Estimator)**.** We have

$$\widehat{\Phi} \xrightarrow{p} \Phi$$

and

$$\sqrt{T}\left(\mathrm{vec}\widehat{\Phi} - \mathrm{vec}\Phi\right) \rightsquigarrow N\left(0, \Gamma_p^{-1} \otimes \Sigma\right)$$

where $\Gamma_p = \mathrm{plim}\frac{1}{T}(X'X)$.

In order to make use of this result in practice, we replace Γ_p by its estimate $\widehat{\Gamma}_p = X'X/T$ and replace Σ by its estimate $\widehat{\Sigma} = \widehat{Z}\widehat{Z}'/T = (Y - \widehat{\Phi}X')(Y - \widehat{\Phi}X')'/T$.

10.2 Yule-Walker Estimation

We can again use the Yule-Walker estimation for VAR models. Consider first a VAR(1) model. In this case the Yule-Walker equation is simply

$$\begin{aligned} \Gamma(0) &= \Phi\Gamma(-1) + \Sigma \\ \Gamma(1) &= \Phi\Gamma(0) \end{aligned}$$

or

$$\begin{aligned}\Gamma(0) &= \Phi\Gamma(0)\Phi' + \Sigma \\ \Gamma(1) &= \Phi\Gamma(0).\end{aligned}$$

The solution is

$$\begin{aligned}\Phi &= \Gamma(1)\Gamma(0)^{-1} \\ \Sigma &= \Gamma(0) - \Gamma(1)\Gamma(0)^{-1}\Gamma(1)'.\end{aligned}$$

Replacing the theoretical moments by their empirical counterparts, we get the Yule-Walker estimator for Φ and Σ:

$$\begin{aligned}\widehat{\Phi} &= \widehat{\Gamma}(1)\widehat{\Gamma}(0)^{-1} \\ \widehat{\Sigma} &= \widehat{\Gamma}(0) - \widehat{\Phi}\widehat{\Gamma}(0)\widehat{\Phi}'.\end{aligned}$$

In general the Yule-Walker estimator for VAR(p) is given by the solution to the system of equations

$$\begin{cases}\widehat{\Gamma}(h) = \sum_{j=1}^{p} \widehat{\Phi}_j \widehat{\Gamma}(h-j), & h = 1, \ldots, p \\ \widehat{\Sigma} = \widehat{\Gamma}(0) - \widehat{\Phi}_1\widehat{\Gamma}(-1) - \cdots - \Phi_p\widehat{\Gamma}(-p).\end{cases}$$

The Yule-Walker estimator and OLS estimator are asymptotically equivalent. In fact, they yield very similar estimates even for finite samples. However, as in the univariate case, the Yule-Walker estimator always delivers, in contrast to the least-square estimator, coefficient estimates with the property $\det(I_n - \hat{\Phi}_1 z - \cdots - \hat{\Phi}_p z^p) \neq 0$ for all $z \in \mathcal{C}$ with $|z| \leq 1$. Thus, the Yule-Walker estimator guarantees that the estimated VAR possesses a causal representation. This, however, comes at the price that the Yule-Walker estimator has a larger small-sample bias than the least-squares estimator, especially when the roots of $\Phi(z)$ get close to the unit circle. Thus, it is generally preferable to use the least-squares estimator in practice.

11 Forecasting with VAR Models

For VAR(1) model, the forecast is

$$\mathbb{P}_T X_{T+h} = \Phi^h X_T.$$

Note that $X_{T+h} = Z_{T+h} + \Phi Z_{T+h-1} + \cdots + \Phi^{h-1} Z_{T+1} + \Phi^h X_T$. So the covariance matrix of the forecast error is

$$\begin{aligned}\mathbb{E}(X_{T+h} - \Phi^h X_T)(X_{T+h} - \Phi^h X_T)' &= \Sigma + \Phi\Sigma\Phi' + \cdots + \Phi^{h-1}\Sigma\Phi'^{h-1} \\ &= \sum_{j=0}^{h-1} \Phi^j \Sigma \Phi'^j.\end{aligned}$$

For example, for $h = 1$, the error is Z_t, so the covariance matrix of the forecast error is Σ. For $h = 2$, the covariance matrix of the forecast error is $\Sigma + \Phi\Sigma\Phi'$, etc.

The forecast for VAR(p) can be computed recursively, as in Section 3.1.1. For example, for $h = 1$, the forecast is

$$\mathbb{P}_T X_{T+1} = \Phi_1 X_T + \Phi_2 X_{T-1} + \cdots + \Phi_p X_{T+1-p}$$

with forecast error $X_{T+1} - \mathbb{P}_T X_{T+1} = Z_t$ which has mean zero and covariance matrix Σ. For $h = 2$ the forecast is

$$\mathbb{P}_T X_{T+2} = (\Phi_1^2 + \Phi_2) X_T + (\Phi_1\Phi_2 + \Phi_3) X_{T-1} + \cdots + (\Phi_1\Phi_{p-1} + \Phi_p) X_{T+2-p} + (\Phi_1\Phi_p) X_{T+1-p}.$$

In general, we can write a VAR(p) process in its causal representation as

$$X_t = Z_t + \Psi_1 Z_{t-1} + \Psi_2 Z_{t-2} + \cdots.$$

Thus

$$X_{T+h} = Z_{T+h} + \Psi_1 Z_{T+h-1} + \Psi_2 Z_{T+h-2} + \cdots + \Psi_h Z_T + \Psi_{h+1} Z_{t-1} + \cdots$$

The forecast for X_{T+h} would be

$$\mathbb{P}_T X_{T+h} = \Psi_h Z_T + \Psi_{h+1} Z_{t-1} + \cdots$$

so that the forecast error is

$$\text{error } = Z_{T+h} + \Psi_1 Z_{T+h-1} + \cdots + \Psi_{h-1} Z_{T+1}.$$

The variance of the forecast error is thus

$$\Sigma + \Psi_1 \Sigma \Psi_1' + \cdots + \Psi_{h-1} \Sigma \Psi_{h-1}' = \sum_{j=0}^{h-1} \Psi_j \Sigma \Psi_j'.$$

11.1 Forecasting with Estimated Parameters

Note that in the above derivations we assumed that the true parameters are known. If we replace the parameters by their estimates, then the forecast errors would increase:

$$\begin{aligned} X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h} &= (X_{T+h} - \mathbb{P}_T X_{T+h}) + \left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right) \\ &= \sum_{j=0}^{h-1} \Psi_j Z_{T+h-j} + \left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right). \end{aligned}$$

The covariance matrix of the forecast errors would then be different as well. For example, for $h = 1$ and $p = 1$, the covariance matrix is

$$\Sigma + \frac{n}{T}\Sigma = \frac{T+n}{T}\Sigma.$$

For $h = 1$ and general p the covariance matrix is

$$\Sigma + \frac{np}{T}\Sigma = \frac{T+np}{T}\Sigma.$$

(This is only an approximation as we applied asymptotic results to small samples)

11.2 Selecting the Order

The selection of order p for VAR models can also be based on information criteria like AIC, BIC and HQC. We omit the formula here.

11.3 Evaluating the Forecasts

The forecasts can be evaluated by:

1. Root-mean-squared-error:
$$\sqrt{\frac{1}{h}\sum_{T+1}^{T+h}(\widehat{X}_{it} - X_{it})^2}.$$
2. Mean-absolute error:
$$\frac{1}{h}\sum_{T+1}^{T+h}|\widehat{X}_{it} - X_{it}|.$$

12 Interpretation and Identification of VAR Models

12.1 Wiener-Granger Causality

Consider $X_t = (X_{1t}, X_{2t})'$. Denote the variance of the forecast error for X_{it} by $v_1(h)$, and denote $\tilde{v}_1(h)$ the variance of the forecast error for X_{it} when X_{2t} is omitted from the model. According to Granger, the second variable X_{2t} *causes* or is *causal* for X_{1t} if

$$v_1(h) < \tilde{v}_1(h) \qquad \text{for some } h \geq 1,$$

i.e. the presence of $\{X_{2t}\}$ helps in improving the forecast for $\{X_{1t}\}$.

In the context of VAR(1) model, the one-period forecast is

$$\mathbb{P}_T X_{T+1} = \begin{pmatrix} \mathbb{P}_T X_{1,T+1} \\ \mathbb{P}_T X_{2,T+1} \end{pmatrix} = \Phi X_T = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} X_{1,T} \\ X_{2,T} \end{pmatrix}$$

and so

$$\mathbb{P}_T X_{1,T+1} = \phi_{11} X_{1T} + \phi_{12} X_{2T}.$$

If $\phi_{12} = 0$ then the second variable does not contribute to the one-period forecast of the first variable. So the test for Granger-causality is a test about whether $\phi_{12} = 0$. In general the one-period forecast for VAR(p) models is

$$\mathbb{P}_T X_{1,T+1} = \phi_{11}^{(1)} X_{1T} + \phi_{12}^{(1)} X_{2T} + \cdots + \phi_{11}^{(p)} X_{1,T+1-p} + \phi_{12}^{(p)} X_{2,T+1-p}$$

and so the test for Granger-causality is about whether $\phi_{12}^{(1)} = \cdots = \phi_{12}^{(p)} = 0$. The hypothesis can be tested using a Wald test (F-test). In the context of a VAR(1) model a simple t-test is also possible.

12.2 Structural and Reduced Form

We follow the notations used in [Neu16]. The notation is meant to be general and inclusive, as it involved two structural matrices A and B.

Structural form of SVAR:

$$AX_t = \Gamma_1 X_{t-1} + \cdots + \Gamma_p X_{t-p} + BV_t$$

where the diagonals of A and B are normalized to 1s (by normalization we mean divide every row by the diagonal element) and $V_t \sim WN(0, \Omega)$ with Ω being a diagonal matrix. The reduced form is

$$\begin{aligned} X_t &= A^{-1}\Gamma_1 X_{t-1} + \cdots + A^{-1}\Gamma_p X_{t-p} + A^{-1} B V_t \\ &= \Phi_1 X_{t-1} + \cdots + \Phi_p X_{t-p} + Z_t. \end{aligned}$$

The relationship between the structural shocks V_t and the reduced form disturbances is gicen by

$$Z_t = A^{-1}BV_t.$$

The covariance matrix of Σ, the one that we can directly estimate, is

$$\Sigma = A^{-1}B\Omega B'A'^{-1}. \tag{11}$$

The identification problem is to infer A, B and Ω from estimation of Σ. The first step is to assume that Ω is diagonal, or is the identity matrix. This corresponds to the assumption that the structural shocks are orthogonal.

Let's take $n = 2$ as example. Assume for simplicity $A = I_2$ (We can always do so. For example we can let $B^* = A^{-1}B$ and then normalize the diagonal of B^* to 1s. Then $Z_t = B^*V_t$ and $\Sigma = B^*\Omega B^{*'}$) We can write out the equation $\Sigma = B\Omega B'$:

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & b_{12} \\ b_{21} & 1 \end{pmatrix} \begin{pmatrix} \omega_1^2 & 0 \\ 0 & \omega_2^2 \end{pmatrix} \begin{pmatrix} 1 & b_{21} \\ b_{12} & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \omega_1^2 + b_{12}^2\omega_2^2 & b_{21}\omega_1^2 + b_{12}\omega_2^2 \\ b_{21}\omega_1^2 + b_{12}\omega_2^2 & b_{21}^2\omega_1^2 + \omega_2^2 \end{pmatrix}$$

This gives the following system of equation

$$\begin{cases} \sigma_1^2 = \omega_1^2 + b_{12}^2\omega_2^2 \\ \sigma_{12} = b_{21}\omega_1^2 + b_{12}^2\omega_2^2 \\ \sigma_2^2 = b_{21}^2\omega_1^2 + \omega_2^2. \end{cases} \tag{12}$$

We have four unknowns $\{b_{12}, b_{21}, \omega_1^2, \omega_2^2\}$ and three equations, so we see from here that we need one additional identification assumption.

12.3 Short-Run Restrictions

Short-run restrictions are essentially restrictions on the coefficients of A and/or B. [Sim80] proposed a simple recursive scheme. The assumption is that B is a lower triangular matrix, so that $b_{12} = 0$. This is compatible with the Cholesky decomposition of Σ. The theorem says that

Let A be a Hermitian positive-definite matrix. Then

$$A = LDL^*$$

where L is lower triangular with ones on the diagonal and D is a diagonal matrix.

As $Z_t = BV_t = \begin{pmatrix} 1 & 0 \\ b_{21} & 1 \end{pmatrix} \begin{pmatrix} V_{1t} \\ V_{2t} \end{pmatrix}$, we have

$$\begin{cases} Z_{1t} = V_{1t} \\ Z_{2t} = b_{21} V_{1t} + V_{2t}. \end{cases}$$

The Interpretation is that V_{1t} is the only structural shock which has an effect on X_{1t} in period t. All other shocks have no contemporaneous effect (think about e.g. a monetary shock cannot have immediate effect on output due to sticky prices) Since $Z_{1t} = V_{1t}$ we have $\sigma_1^2 = \omega_1^2$ so that we solved ω_1^2. The second variable X_{2t} is affected by $\{V_{1t}, V_{2t}\}$ (if $n > 2$ then we are saying X_{2t} is only affected by the two shocks and not by $V_{3t}, \ldots, V_{nt}$). We have $Z_{2t} = b_{21} V_{1t} + V_{2t}$ so $\sigma_{21} = Cov(Z_{2t}, Z_{1t}) = b_{21}\omega_1^2 + 0 = b_{21}\omega_1^2$ (note the orthogonality assumption also enters here) From this we can derive b_{21} and then ω_2^2 from $\sigma_2^2 = Cov(Z_{2t}, Z_{2t}) = \omega_2^2 + b_{21}^2 \omega_1^2$.

12.4 Interpretation of SVAR Models

12.4.1 Impulse Response Functions

After we have identified the model, it is now possible to compute the impulse response functions of the structural shocks. Recall $Z_t = A^{-1}B$ and so

$$\begin{aligned} X_t &= Z_t + \Psi_1 Z_{t-1} + \Psi_2 Z_{t-2} + \cdots \\ &= A^{-1}BV_t + \Psi_1 A^{-1} B V_{t-1} + \Psi_2 A^{-1} B V_{t-2} + \cdots \end{aligned}$$

And so the effect of the j-th structural disturbance ($V_{j,t}$) on the i-th variable ($X_{i,t}$) after h periods is the (i, j) element of the matrix $\Psi_h A^{-1} B$ (i is the row, manifesting to which variable in X_t we address to, and j points to V_j):

$$\frac{\partial X_{i,t+h}}{\partial V_{j,t}} = \left[\Psi_h A^{-1} B\right]_{ij}.$$

12.4.2 Forecast Error Variance Decomposition (FEVD)

After we have identified the model, it is also possible to calculate how much forecast variances that each structural shock contribute to. A concrete statement is: we get error in forecasting X_{it} (for example this variable can be output), and the variance of the error is $v(h)$. How much is the variance of the first structural shock V_{1t} (namely ω_1^2) contribute to $v(h)$? How much is the variance of the second structural shock V_{2t} (namely ω_2^2) contribute to $v(h)$? etc. This gives us a hint as to which shock gives rise to the most *unexpected* movements of the variable of interest.

Recall that the covariance matrix of the forecast error is

$$MSE(h) = \sum_{j=0}^{h-1} \Psi_j \Sigma \Psi_j'.$$

Since by Eq. (11) $\Sigma = A^{-1}B\Omega B'A'^{-1}$, we have

$$\sum_{j=0}^{h-1} \Psi_j \Sigma \Psi_j' = \sum_{j=0}^{h-1} \Psi_j A^{-1} B \Omega B' A'^{-1} \Psi_j'.$$

For example, for $h = 1$, and assume $A = I$ and $\Omega = I$, the equation is

$$\Sigma = BB'.$$

The diagonal of $MSE(h)$ is the variances of the forecast error for variables $X_{1t}, \ldots, X_{nt}$. Expanding the right, each diagonal element of $MSE(h)$ should be a linear combination of $\{\omega_1^2, \ldots, \omega_n^2\}$.

$$m_{ii}^{(h)} = d_{i1}^{(h)}\omega_1^2 + \cdots + d_{in}^{(h)}\omega_n^2$$

where $m_{ii}^{(h)}$ is the i-th diagonal element of $MSE(h)$. Now divide $m_{ii}^{(h)} = \mathbb{V}(\text{forecast error for } X_i)$ by each component in the sum. We then get what is the proportion of $\mathbb{V}(V_{jt}) = \omega_j^2$ in the error variance for each structural shock $j = 1, \ldots, n$.

12.5 Long-Run Restrictions

Instead of restricting some coefficients of A and/or B to be zero, we can use long-run restrictions. Long-run restrictions constrain the long-run effect of structural shocks. Note that the technique only makes sense if some integrated variables are involved, because in the stationary case the effects of all shocks vanish eventually.

The technique was first proposed by [BQ89] and we replicate the example in the paper. They analyzed a two-variable system consisting of logged real GDP denoted by $\{Y_t\}$ and the unemployment rate $\{U_t\}$. Logged GDP is typically integrated of order one, whereas $\{U_t\}$ is considered to be stationary. Thus we can apply the VAR approach to the stationary process $\{X_t\} = \{(\Delta Y_t, U_t)'\}$. Assuming that $\{X_t\}$ is already demeaned and follows a causal VAR process, we have the following representations

$$\begin{aligned} X_t = \begin{pmatrix} \Delta Y_t \\ U_t \end{pmatrix} &= \Phi_1 X_{t-1} + \cdots + \Phi_p X_{t-p} + Z_t \\ &= \Psi(\mathrm{L}) Z_t = Z_t + \Psi_1 Z_{t-1} + \Psi_2 Z_{t-2} + \cdots \end{aligned}$$

For simplicity we assume $A = I_2$ so that

$$Z_t = BV_t = \begin{pmatrix} 1 & b_{12} \\ b_{21} & 1 \end{pmatrix} \begin{pmatrix} v_{dt} \\ v_{st} \end{pmatrix}$$

where $V_t = (v_{dt}, v_{st})' \sim WN(0, \Omega)$ with $\Omega = \mathrm{diag}(\omega_d^2, \omega_d^2)$. Thereby $\{v_{dt}\}$ and $\{v_{st}\}$ denote demand and supply shocks respectively. A demand shock in period t on GDP growth in period $t + h$ is

$$\frac{\partial \Delta Y_{t+h}}{\partial v_{dt}} = [\Psi_h B]_{11}.$$

As Y_{t+h} can be written as $Y_{t+h} = \Delta Y_{t+h} + \Delta Y_{t+h-1} + \cdots + \Delta Y_{t+1} + Y_t$, the effect of the demand shock on the *level* of logged GDP is

$$\frac{\partial Y_{t+h}}{\partial v_{dt}} = \sum_{j=0}^{h} [\Psi_j B]_{11} = \left[\sum_{j=0}^{h} \Psi_j B_{11} \right]_{11}.$$

[BQ89] propose that the long-run effect of the demand shock on the level of logged GDP should be zero:

$$\lim_{h\to\infty} \frac{\partial Y_{t+h}}{\partial v_{dt}} = \sum_{j=0}^{\infty} [\Psi_j B]_{11} = 0.$$

This implies

$$\sum_{j=0}^{\infty} \Psi_j B = \left(\sum_{j=0}^{\infty} \Psi_j \right) B = \Psi(1) \begin{pmatrix} 1 & b_{12} \\ b_{21} & 1 \end{pmatrix} = \begin{pmatrix} 0 & * \\ * & * \end{pmatrix}.$$

This restriction is sufficient to infer b_{21} from the relation $[\Psi(1)]_{11} \cdot 1 + b_{21}[\Psi(1)]_{12} = 0$:

$$b_{21} = -\frac{[\Psi(1)]_{11}}{[\Psi(1)]_{12}} = -\frac{[\Phi(1)^{-1}]_{11}}{[\Phi(1)^{-1}]_{12}}.$$

The long-run effect of the supply shock is left unrestricted and is therefore nonzero in general. Note that b_{21} depends on $\Phi(1)$, and thus on $\Phi_1, \ldots, \Phi_p$. The results are therefore more sensitive to specifications of the VAR model.

After we have derived b_{21}, we have three unknowns $\{b_{12}, \omega_d^2, \omega_s^2\}$ and three equations (Eq. (12)) so we can then solve for the unknowns.

12.5.1 The General Approach

The general case of long-run restrictions has a structure similar to the case of short-run restrictions. Recall our structural model

$$AX_t = \Gamma_1 X_{t-1} + \cdots + \Gamma_p X_{t-p} + BV_t, \qquad V_t \sim WN(0, \Omega).$$

We can write $A(L)X_t = BV_t$ where $A(L) = A - \Gamma_1 L - \cdots - \Gamma_p L^p$. The reduced form is $\Phi(\mathrm{L})X_t = Z_t, Z_t \sim WN(0, \Sigma)$. As can be seen from the relation $X_t = \Phi(\mathrm{L})^{-1} Z_t$, the long-run variance is

$$J = \mathbb{V}(X_t) = \Phi(1)^{-1} \Sigma \Phi(1)^{-1'} = \Psi(1) \Sigma \Psi(1)'. \tag{13}$$

We can the substitute Σ by $A^{-1} B \Omega B' A'^{-1}$ in Eq. (13) to get

$$J = \Big(\Phi(1)^{-1} A^{-1} B \Big) \Omega \Big(B' A'^{-1} \Phi(1)^{-1'} \Big) = \Big(\Psi(1) A^{-1} B \Big) \Omega \Big(B' A'^{-1} \Psi(1)' \Big).$$

We'd like to identify A, B and Ω from the estimation $\widehat{J}$. The identification is through zero restrictions on some elements of $\Psi(1)A^{-1}B$, or $\Phi(1)^{-1}A^{-1}B$. Setting the ij-th element $[\Psi(1)A^{-1}B]_{ij}$ to zero amounts to set the cumulative effect of the j-th disturbance of the j-th structural disturbance $V_{j,t}$ on the i-th variable equal to zero. If the i-th variable enters X_t in first differences, as was the case in [BQ89], this zero restriction restrains the long-run effect on the level of that variable.

One case is to assume $A = I_n$ and $\Psi(1)B$ is lower triangular. In this case B and Ω can be estimated from the Cholesky decomposition of the estimated long-run variance $\widehat{J}$. Let $\widehat{J} = \hat{L}\hat{D}\hat{L}'$ be the Cholesky decomposition. We can write this as

$$\begin{aligned}\widehat{J} &= \hat{L}\hat{D}\hat{L}' \\ &= \hat{\Phi}(1)^{-1}\left(\hat{\Phi}(1)\hat{L}\hat{U}^{-1}\right)\left(\hat{U}\,\hat{D}\hat{U}\right)\left(\hat{U}^{-1}\hat{L}'\hat{\Phi}(1)'\right)\hat{\Phi}(1)^{-1'}\end{aligned}$$

As $\widehat{J} = \hat{\Phi}(1)^{-1}B\hat{\Omega}B'\hat{\Phi}(1)^{-1'}$ we can estimate B as $\left(\hat{\Phi}(1)\hat{L}\hat{U}^{-1}\right)$ and Ω as $\left(\hat{U}\,\hat{D}\hat{U}\right)$. The matrix $U = \text{diag}(\hat{\Phi}(1)\hat{L})$ is used to assure the normalization of the diagonal of $\hat{B}$ to 1.

In the above we used the method of moments approach, i.e. we used the long-run variance J (second moment) to estimate the parameters. Another approach is instrumental variable (IV). As for this example, we write the reduced form as

$$\Delta X_t = -\Phi(1)X_{t-1} + \widetilde{\Phi}_1\Delta X_{t-1} + \cdots + \widetilde{\Phi}_{p-1}\Delta X_{t-p+1} + Z_t,$$

where $\widetilde{\Phi}_j = -\sum_{i=j+1}^{p}\Phi_i$, $j - 1, 2, \ldots, p-1$. Assume $B = I_n$ so that $AZ_t = V_t$. Multiply the above equation by A:

$$A\Delta X_t = -A\Phi(1)X_{t-1} + A\widetilde{\Phi}_1\Delta X_{t-1} + \cdots + A\widetilde{\Phi}_{p-1}\Delta X_{t-p+1} + V_t. \tag{14}$$

Now assume $A\Phi(1)$ is *lower* triangular, so that $[A\Phi(1)]^{-1}$ is *upper* triangular. Recall that the long-run variance in this case is

$$\begin{aligned}J &= \left(\Phi(1)^{-1}A^{-1}\right)\Omega\left(\Phi(1)^{-1}A^{-1}\right)' \\ &= \Big[A\Phi(1)\Big]^{-1}\Omega\Big[A\Phi(1)\Big]^{-1'}\end{aligned}$$

and so this amounts to saying that the structural shocks $\{V_{2t}, \ldots, V_{nt}\}$ have no long-run impact on the first variable X_{1t}. It is therefore possible to estimate the coefficients $\{A_{12}, \ldots, A_{1n}\}$ using $X_{2,t-1}, \ldots, X_{n,t-1}$ as instruments.

For the $n = 2$ case Eq. (14) is

$$\begin{pmatrix}1 & A_{12}\\ A_{21} & 1\end{pmatrix}\begin{pmatrix}\Delta\tilde{X}_{1t}\\ \Delta\tilde{X}_{2t}\end{pmatrix} = -\begin{pmatrix}[A\Phi(1)]_{11} & 0\\ [A\Phi(1)]_{21} & [A\Phi(1)]_{22}\end{pmatrix}\begin{pmatrix}\tilde{X}_{1,t-1}\\ \tilde{X}_{2,t-1}\end{pmatrix} + \begin{pmatrix}V_{1t}\\ V_{2t}\end{pmatrix}.$$

Multiply this out:

$$\begin{cases} \Delta\tilde{X}_{1t} = -A_{12}\Delta\tilde{X}_{2t} - [A\Phi(1)]_{11}\tilde{X}_{1,t-1} + V_{1t} \\ \Delta\tilde{X}_{2t} = -A_{21}\Delta\tilde{X}_{1t} - [A\Phi(1)]_{21}\tilde{X}_{1,t-1} - [A\Phi(1)]_{22}\tilde{X}_{2,t-1} + V_{2t}. \end{cases}$$

Thereby $\Delta\tilde{X}_{1t}$ and $\Delta\tilde{X}_{2t}$ denote the OLS residuals from a regression of ΔX_{1t} and ΔX_{2t} on

$$\left\{\Delta X_{1,t-1}, \Delta X_{2,t-1}, \ldots, \Delta X_{1,t-p+1}, \Delta X_{2,t-p+1}\right\}.$$

We see that $\tilde{X}_{2,t-1}$ is a valid instrument for $\Delta\tilde{X}_{2t}$ (relevance is from the second equation, and exclusion restriction is from the fact that it does not appear in the first equation). Thus we can estimate A_{12} using the IV approach. For the estimation of A_{21}, we can use the residuals from the first equation as instruments because V_{1t} and V_{2t} are assumed to be uncorrelated.

13 Appendix

- vec operator: stack columns of a matrix A on top of each other:

$$\text{vec}(A) = [a_{11}, \ldots, a_{m1}, a_{12}, \ldots, a_{m2}, \ldots, a_{1n}, \ldots, a_{mn}]^T.$$

  For example, for $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, the vectorization is $\text{vec}(A) = \begin{pmatrix} a \\ c \\ b \\ d \end{pmatrix}$.

- Kronecker product: if A is an $m \times n$ matrix and B is a $p \times q$ matrix, then the Kronecker product $A \otimes B$ is the $mp \times nq$ block matrix:

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$

  Example:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \otimes \begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix} = \begin{pmatrix} 1 \cdot \begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix} & 2 \cdot \begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix} \\ 3 \cdot \begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix} & 4 \cdot \begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 1 \cdot 0 & 1 \cdot 5 & 2 \cdot 0 & 2 \cdot 5 \\ 1 \cdot 6 & 1 \cdot 7 & 2 \cdot 6 & 2 \cdot 7 \\ 3 \cdot 0 & 3 \cdot 5 & 4 \cdot 0 & 4 \cdot 5 \\ 3 \cdot 6 & 3 \cdot 7 & 4 \cdot 6 & 4 \cdot 7 \end{pmatrix} = \begin{pmatrix} 0 & 5 & 0 & 10 \\ 6 & 7 & 12 & 14 \\ 0 & 15 & 0 & 20 \\ 18 & 21 & 24 & 28 \end{pmatrix}.$$

- Property:

$$\text{vec}(ABC) = (C^T \otimes A)\text{vec}(B).$$

References

Blanchard, O. J. and D. Quah (Sept. 1989). “The Dynamic Effects of Aggregate Demand and Supply Disturbances”. In: *American Economic Review* 79.4, pp. 655–673.

Neusser, K. (Dec. 2016). *Time Series Econometrics*. Springer Texts in Business and Economics 978-3-319-32862-1. Springer.

Sims, C. A. (Jan. 1980). “Macroeconomics and Reality”. In: *Econometrica* 48.1, pp. 1–48.