Chapter 1: Time Series and Their Features

1.1 Basic Definitions

1.1.1 Observation period

A time series on some variable x will be denoted as x_t, where the subscript t represents time, with $t = 1$ being the first observation available on x and $t = T$ being the last. The complete set of times $t = 1; 2; \ldots; T$ will often be referred to as the observation period. *The observations are typically measured at equally spaced intervals.*

1.1.2 Forecast horizon

Unknown values of x_t at, say, times $T + 1; T + 2; \ldots; T + h$, where h is referred to as the *forecast horizon.*

1.1.3 Autocorrelations

Definition of $lag - k\ autocorrelation$

$$r_k = \frac{\sum_{t=k+1}^{T}(x_t - \bar{x})(x_{t-k} - \bar{x})}{Ts^2}$$

where $\bar{x} = T^{-1}\sum_{t-1}^{T} x_t$, and $s^2 = T^{-1}\sum_{t=1}^{T}(x_t - \bar{x})$ are the sample mean and variance of x_t.

The set of sample autocorrelations for various values of k is known as the *sample autocorrelation function* (SACF) and plays a key role in time series analysis. *When a time series is observed at monthly or quarterly intervals an annual seasonal pattern is often an important feature.*

1.1.4 Stationarity and Nonstationarity

Typical condition to judge is whether the series has a **constant mean level or variance** .

1.1.5 Trends

Including *linear trends* and *non-linear trends*, depending on whether the trend has **constant slopes**.

1.1.6 Counts

Some series occur naturally as (small) integers and these are often referred to as *counts*.

1.2 Notices

- **Common Features:** Two or more time series may contain common features. Two series which share a common trend and are said to *cointegrate*.

- **Natural Constraints:** Some time series have natural constraints placed upon them. Such compositional time series require distinctive treatment through special transformations before they can be analyzed.

Chapter 2: Transforming Time Series

2.1 Transformations

2.1.1 Distributional Transformations

Raw distributions need to be transformed to exhibit some distributional properties (such as normal distribution).

> For example, if a series is only able to take positive (or at least nonnegative) values, then its distribution will usually be skewed to the right, because although there is a natural lower bound to the data, often zero, no upper bound exists and the values are able to "stretch out," possibly to infinity. In this case a simple and popular transformation is to take logarithms, usually to the base e (natural logarithms).

The ratio of cumulative standard deviations:

$$s_i(x)/s_i(logx),\ where\ s_i^2(x) = i^{-1}\sum_{t=1}^{i}(x_t-\bar{x_i})^2,\ \bar{x_i} = i^{-1}\sum_{t=1}^{i}x_t$$

Since this ratio increases monotonically throughout the observation period, the logarithmic transformation clearly helps to stabilize the variance whenever the standard deviation of a series is proportional to its level.

The availability of a more general class of transformations would be useful for attaining approximate normality.

- **Box-Cox Transformation**

  A class of *power transformations* that contains the logarithmic as a special case is that proposed by Box and Cox (1964) for positive x:

$$f^{\mathrm{BC}}(x_t,\lambda) = \begin{cases}\left(x_t^\lambda - 1\right)/\lambda & \lambda \neq 0 \\ \log x_t & \lambda = 0\end{cases}$$

  The restriction to positive values that is required by the Box-Cox transformation can be relaxed in several ways. A shift parameter may be introduced above to handle situations where x may take negative values but is still bounded below, but this may lead to inferential problems when λ is estimated.

- **Signed Power Transformation**

  A possible alternatives are the *signed power transformation* proposed by Bickel and Doksum:

$$f^{SP}(x_t,\lambda) = (sgn(x_t)|x_t^\lambda| - 1)/\lambda\ , \lambda > 0$$

- **Generalized Power (GP) Transformation**

$$f^{\mathrm{GP}}\left(x_t, \lambda\right)=\begin{cases}\left(\left(x_t+1\right)^\lambda-1\right) / \lambda & x_t \geq 0, \lambda \neq 0 \\ \log \left(x_t+1\right) & x_t \geq 0, \lambda=0 \\ -\left(\left(-x_t+1\right)^{2-\lambda}-1\right) /(2-\lambda) & x_t<0, \lambda \neq 2 \\ -\log \left(-x_t+1\right) & x_t<0, \lambda \neq 2\end{cases}$$

- **Inverse Hyperbolic Sine (IHS) Transformation**

$$f^{\mathrm{IHS}}\left(x_t, \lambda\right)=\frac{\sinh ^{-1}\left(\lambda x_t\right)}{\lambda}=\log \frac{\lambda x_t+\left(\lambda^2 x_t^2+1\right)^{1 / 2}}{\lambda} \quad \lambda>0$$

The transformation parameter λ may be estimated by the method of **maximum likelihood (ML)**.

> Suppose that for a general transformation $f(x_t, \lambda)$, the model $f(x_t, \lambda) = \mu_t + a_t$ at is assumed, where μ_t is a model for the mean of $f(x_t, \lambda)$ and a_t is assumed to be independent and normally distributed with zero mean and constant variance. The ML estimator λ^ is then obtained by maximizing over λ the concentrated log-likelihood function:
>
> $$\ell(\lambda)=C_f-\left(\frac{T}{2}\right) \sum_{t=1}^T \log \hat{a}_t^2+D_f\left(x_t, \lambda\right)$$
>
> where $\hat{a}_t=f\left(x_t, \lambda\right)-\hat{\mu}_t$ are the residuals from ML estimation of the model, C_f is a constant and $D_f(x_t, \lambda)$ depends on which of the transformations (2.1) - (2.4) is being used:
>
> $$\begin{aligned} D_f\left(x_t, \lambda\right) & =(\lambda-1) \sum_{t=1}^T \log \left|x_t\right| & & \text{for BC and SP Transformation} \\ & =(\lambda-1) \sum_{t=1}^T \operatorname{sgn}\left(x_t\right) \log \left(\left|x_t\right|+1\right) & & \text{for GP Transformation} \\ & =-\frac{1}{2} \sum_{t=1}^T \log \left(1+\lambda^2 x_t^2\right) & & \text{for IHS Transformation}\end{aligned}$$
>
> If $\hat{\lambda}$ is the ML estimator and $\ell(\hat{\lambda})$ is the accompanying maximized likelihood from above, then a confidence interval for λ can be constructed using the standard result that $2(\ell(\hat{\lambda})-\ell(\lambda))$ is asymptotically distributed as $\chi^2(1)$, so that a 95% confidence interval, for example, is given by those values of λ for which $\ell(\hat{\lambda})-\ell(\lambda)<1.92$.

2.1.2 Stationarity Including Transformations

A simple stationarity transformation is to take successive differences of a series, on defining the *first-difference* of x_t as $\nabla x_t = x_t - x_{t-1}$.

It can eradicate the trends in both series and differencing has a lot to recommend it both practically and theoretically for transforming a nonstationary series to stationarity.

Some caution is required when taking higher-order differences that $\nabla^2 x_t = (1-B)^2 x_t = \left(1-2B+B^2\right)x_t = x_t - 2x_{t-1} + x_{t-2}$ is not equal to $x_t - x_{t-2} = \nabla_2 x_t$, where *lag operator B*:

$$B^i x_t \equiv x_{t-j}$$

and the notation

$$\nabla_i = 1 - B^j$$

for the taking of j-period differences is defined.

2.1.3 Decomposing a Time Series and Smoothing Transformations

It is often the case that the long-run behavior of a time series is of particular interest and attention is then focused on isolating these long-term movements from shorter-run by separating the observations through a decomposition, generally of the form “data = fit + residual.” Because such a decomposition is more than likely going to lead to a smooth series, this might be better thought of as “data = smooth + rough,” terminology borrowed from Tukey.

> **Moving Averages(MA):** The simplest MA replaces x_t with the average of itself, its predecessor, and its successor, that is, by the $\mathbf{MA}(3)\frac{1}{3}(x_{t-1}+x_t+x_{t+1})$. More complicated formulations are obviously possible: the (2n+1)-term weighted and centered MA [WMA(2n+1)] replaces x_t with
>
> $$\mathrm{WMA}_t(2n+1) = \sum_{i=-n}^{n} \omega_i x_{t-i}$$
>
> where the *weights* ω_i are restricted to sum to unity: $\sum_{i=-n}^{n} \omega_i = 1$, and are often symmetrically valued about the central weight ω_0.

As more terms are included in the MA, the smoother it becomes, albeit with the trade-off that since n observations are “lost” at the beginning and at the end of the sample, more observations will be lost the larger n is. If observations at the end of the sample, the most “recent”, are more important than those at the start of the sample, then an uncentered MA may be considered, such as $\sum_{i=0}^{n} \omega_i x_{t-i}$.

The MAs tend to be interpreted as trends; the long-run, smoothly evolving component of a time series, or to say, the “smooth” of a two-component decomposition.

> When a time series is observed at a frequency greater than annual, say monthly or quarterly, a three-component decomposition is often warranted, with the observed series, now denoted Xt, being decomposed into trend, Tt, seasonal, St, and irregular, It, components. The decomposition can either be additive:
>
> $$X_t = T_t + S_t + I_t$$
>
> or multiplicative:
>
> $$X_t = T_t \times S_t \times I_t$$
>
> although the distinction is to some extent artificial, as taking logarithms of multiplication will produce an additive decomposition for $\log X_t$.The seasonal component is a regular, short-term, annual cycle, while the irregular component is what is left over after the trend and seasonal components have been removed; it should with thus be random and hence unpredictable.

> The seasonally adjusted series is then defined as either:
>
> $$X_t^{\mathrm{SA,A}}=X_t-S_t=T_t+I_t$$
>
> or
>
> $$X_t^{\mathrm{SA,M}}=\frac{X_t}{S_t}=T_t\times I_t$$
>
> depending on which form of decomposition is used.

2.2 Notices

- For some time series, interpretation can be made easier by taking *proportional or percentage* changes rather than simple differences, that is, transforming by $\nabla x_t/x_{t-1}$ or $100\nabla x_t/x_{t-1}$. For financial time series these are typically referred to as the *return*. When attention is focused on the percentage change in a price index, then these changes are typically referred to as the *rate of inflation*.
- Relationship between the rate of change of a variable and its logarithm:

$$\frac{x_t-x_{t-1}}{x_{t-1}}=\frac{x_t}{x_{t-1}}-1\approx\log\frac{x_t}{x_{t-1}}=\log x_t-\log x_{t-1}=\nabla\log x_t$$

- where the approximation follows from the fact that $\log(1+y)\approx y$ for small y.

Chapter 3: ARMA Models for Stationary Time Series

3.1 Stochastic Processes and Stationarity

Specifying the complete form of the probability distribution, however, will typically be too ambitious a task, so attention is usually concentrated on the first and second moments; the T means:

$$E\left(x_1\right),E\left(x_2\right),\ldots,E\left(x_T\right)$$

T variances:

$$V\left(x_1\right),V\left(x_2\right),\ldots,V\left(x_T\right)$$

and $T(T-1)/2$covariances:

$$\operatorname{Cov}\left(x_i,x_j\right),\quad i<j$$

If the distribution could be assumed to be (multivariate) normal, then this set of expectations would completely characterize the properties of the stochastic process.

It should be emphasized that the procedure of using a single realization to infer the unknown parameters of a joint probability distribution is only valid if the process is ergodic, which roughly means that the sample moments for finite stretches of the realization approach their population counterparts as the length of the realization becomes infinite. Since it is difficult to test for ergodicity using just (part of) a single realization, it will be assumed that this property holds for every time series.

If a series is stationary, then its means and variances should be constant, which means:

$$E\left(x_1\right)=E\left(x_2\right)=\cdots=E\left(x_T\right)=\mu$$

and

$$V\left(x_1\right)=V\left(x_2\right)=\cdots=V\left(x_T\right)=\sigma_x^2$$

This leads to the definition of the *lag-k autocovariance* as:

$$\gamma_k=\operatorname{Cov}\left(x_t, x_{t-k}\right)=E\left(\left(x_t-\mu\right)\left(x_{t-k}-\mu\right)\right), \quad \gamma_0=E(x_t-\mu)^2=V\left(x_t\right)=\sigma_x^2$$

and the *lag-k autocorrelation* may then be defined as

$$\rho_k=\frac{\operatorname{Cov}\left(x_t, x_{t-k}\right)}{\left(V\left(x_t\right) V\left(x_{t-k}\right)\right)^{1/2}}=\frac{\gamma_k}{\gamma_0}=\frac{\gamma_k}{\sigma_x^2}$$

The set of assumptions that the mean and variance of xt are both constant and the autocovariances and autocorrelations depend only on the lag k is known as *weak or covariance stationarity*.

Weak stationarity and strict stationarity: If joint normality could be assumed so that the distribution was entirely characterized by the first two moments, weak stationarity would indeed imply strict stationarity.

Autocorrelation function (ACF): The set of autocorrelations ρ_k, when considered as a function of k, is referred to as the (population) autocorrelation function (ACF).

$$\gamma_k=\operatorname{Cov}\left(x_t, x_{t-k}\right)=\operatorname{Cov}\left(x_{t-k}, x_t\right)=\operatorname{Cov}\left(x_t, x_{t+k}\right)=\gamma_{-k}$$

it follows that $\rho_{-k}=\rho_k$ and so only the positive half of the ACF is usually given.

> The ACF plays a major role in modeling dependencies between the values of x_t since it characterizes, along with the process mean $\mu=E\left(x_t\right)$ and variance $\sigma_x^2=\gamma_0=V\left(x_t\right)$, the stationary stochastic process describing the evolution of xt. It therefore indicates, by measuring the extent to which one value of the process is correlated with previous values, the length and strength of the “memory” of the process.

3.2 Wold’s Decomposition

Theorem: every weakly stationary, purely nondeterministic (any deterministic components have been subtracted from $x_t-\mu$), stochastic process $x_t-\mu$ can be written as a linear combination (or linear filter) of a sequence of uncorrelated random variables.

The linear filter:

$$x_t-\mu=a_t+\psi_1 a_{t-1}+\psi_2 a_{t-2}+\cdots=\sum_{j=0}^{\infty} \psi_j a_{t-j} \quad \psi_0=1$$

The $a_t, t=0, \pm 1, \pm 2, \ldots$ are a sequence of uncorrelated random variables, often known as *innovations*, drawn from a fixed distribution with:

$$E\left(a_t\right)=0 \quad V\left(a_t\right)=E\left(a_t^2\right)=\sigma^2<\infty$$

and

$$\operatorname{Cov}\left(a_t, a_{t-k}\right)=E\left(a_t a_{t-k}\right)=0, \quad \text{for all } k \neq 0$$

Such a sequence is known as a white noise process, and occasionally the innovations will be denoted as $a_t \sim \mathrm{WN}\left(0, \sigma^2\right)$ The coefficients (possibly infinite in number) in the linear filter are known as *ψ-weights*.

Autocorrelation: It is easy to show that the model (3.2) leads to autocorrelation in xt. From this equation it follows that:

$$E\left(x_t\right)=\mu$$

and

$$\begin{aligned}
\gamma_0 &= V\left(x_t\right)=E(x_t-\mu)^2 \\
&= E(a_t+\psi_1 a_{t-1}+\psi_2 a_{t-2}+\cdots)^2 \\
&= E\left(a_t^2\right)+\psi_1^2 E\left(a_{t-1}^2\right)+\psi_2^2 E\left(a_{t-2}^2\right)+\cdots \\
&= \sigma^2+\psi_1^2 \sigma^2+\psi_2^2 \sigma^2+\cdots \\
&= \sigma^2 \sum_{j=0}^{\infty} \psi_j^2
\end{aligned}$$

by using the white noise result that $E\left(a_{t-i} a_{t-j}\right)=0 \text{ for } i \neq j$. Now:

$$\begin{aligned}
\gamma_k &= E\left(x_t-\mu\right)\left(x_{t-k}-\mu\right) \\
&= E\left(a_t+\psi_1 a_{t-1}+\cdots+\psi_k a_{t-k}+\cdots\right)\left(a_{t-k}+\psi_1 a_{t-k-1}+\cdots\right) \\
&= \sigma^2\left(1 \cdot \psi_k+\psi_1 \psi_{k+1}+\psi_2 \psi_{k+2}+\cdots\right) \\
&= \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k}
\end{aligned}$$

and this implies

$$\rho_k=\frac{\sum_{j=0}^{\infty} \psi_j \psi_{j+k}}{\sum_{j=0}^{\infty} \psi_j^2}$$

Converge: If the number of *ψ-weights* in the filter is infinite, the weights must be assumed to be absolutely summable, so that $\sum_{j=0}^{\infty}\left|\psi_j\right|<\infty$, in which case the linear filter representation is said to *converge*. This condition can be shown to be equivalent to assuming that x_t is stationary, and guarantees that all moments exist and are independent of time, in particular that the variance of x_t, γ_0, is finite.

3.3 First-order Autoregressive Process

Definition: Taking $\mu=0$ without loss of generality, choosing $\psi_j=\phi^j$ allows the linear filter to be written as:

$$\begin{aligned}
x_t &= a_t+\phi a_{t-1}+\phi^2 a_{t-2}+\cdots \\
&= a_t+\phi\left(a_{t-1}+\phi a_{t-2}+\cdots\right) \\
&= \phi x_{t-1}+a_t
\end{aligned}$$

or

$$x_t-\phi x_{t-1}=a_t$$

This is known as a *first-order autoregressive process*, often given the acronym AR(1).

The lag operator B allows (possibly infinite) lag expressions to be written in a concise way. For example, by using this operator the AR(1) process can be written as:

$$(1-\phi B)x_t = a_t$$

so that

$$\begin{aligned} x_t &= (1-\phi B)^{-1}a_t = \left(1+\phi B+\phi^2 B^2+\cdots\right)a_t \\ &= a_t+\phi a_{t-1}+\phi^2 a_{t-2}+\cdots \end{aligned}$$

This linear filter representation will converge if $|\phi| < 1$, which is, therefore, the *stationarity condition*.

3.4 First-order Moving Average Process

Definition: Consider the model obtained by choosing $\psi_1 = -\theta \text{ and } \psi_j = 0,\ j \geq 2$, in linear filter:

$$x_t = a_t - \theta a_{t-1} = (1-\theta B)a_t$$

This is known as the *first-order moving average (MA(1))* process and it follows immediately that:

$$\gamma_0 = \sigma^2\left(1+\theta^2\right) \quad \gamma_1 = -\sigma^2\theta \quad \gamma_k = 0 \text{ for } k > 1$$

hence, its ACF is described by

$$\rho_1 = -\frac{\theta}{1+\theta^2} \quad \rho_k = 0 \text{ for } k > 1$$

Thus, although observations one period apart are correlated, observations more than one period apart are not.

3.5 General AR and MA Process

The general **autoregressive** model of order p (AR(p)) can be written as:

$$x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \cdots - \phi_p x_{t-p} = a_t$$

or

$$\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)x_t = \phi(B)x_t = a_t$$

The linear filter representation $x_t = \phi^1(B)a_t = \psi(B)a_t$ can be obtained by equating coefficients in $\phi(B)\psi(B) = 1$.

The general **MA** of order q (MA(q)) can be written as:

$$x_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

or

$$x_t = (1-\theta_1 B-\cdots-\theta_q B^q)a_t = \theta(B)a_t$$

The ACF can be shown to be:

$$\begin{gathered} \rho_k = \frac{-\theta_k+\theta_1\theta_{k+1}+\cdots+\theta_{q-k}\theta_q}{1+\theta_1^2+\cdots+\theta_q^2} \quad k = 1,2,\ldots,q \\ \rho_k = 0 \quad k > q \end{gathered}$$

The ACF of an MA(q) process therefore cuts off after lag q; the memory of the process extends q periods, observations more than q periods apart being uncorrelated.

3.6 Autoregressive-Moving Average Models

First-order autoregressive-moving average (ARMA(1,1)):

$$x_t - \phi x_{t-1} = a_t - \theta a_{t-1}$$

or

$$(1 - \phi B) x_t = (1 - \theta B) a_t$$

The resultant ARMA(p,q) process has the stationarity and invertibility conditions associated with the constituent AR(p) and MA(q) processes respectively.

3.7 Notices

- The expression for ρ_1 can be written as the quadratic equation $\rho_1 \theta^2 + \theta + \rho_1 = 0$. Since θ must be real, it follows that $|\rho_1| < 0.5$.However, both θ and $1/\theta$ will satisfy this equation, and thus, two MA(1) processes can always be found that correspond to the same ACF.

Chapter 4: ARIMA Models for Nonstationary Time Series

4.1 Nonstationarity

To deal with nonstationarity, we begin by characterizing a time series as the sum of a nonconstant mean level plus a random error component:

$$x_t = \mu_t + \varepsilon_t$$

The nonconstant mean level μ_t can be modeled in a variety of ways. One potentially realistic possibility is that the mean evolves as a (nonstochastic) polynomial of order d in time, with the error ε_t assumed to be a stochastic, stationary, but possibly autocorrelated, zero mean process. Thus,

$$x_t = \mu_t + \varepsilon_t = \sum_{j=0}^{d} \beta_j t^j + \psi(B) a_t, \quad E(\varepsilon_t) = \psi(B) E(a_t) = 0$$

we have

$$E(x_t) = E(\mu_t) = \sum_{j=0}^{d} \beta_j t^j$$

and, as the β_j coefficients remain constant through time, such a trend in the mean is said to be *deterministic.*

4.2 ARIMA Processes

If autocorrelation is modeled by an ARMA(p,q) process, then the model for the original series is of the form:

$$\phi(B) \nabla^d x_t = \theta_0 + \theta(B) a_t$$

which is said to be an *autoregressive-integrated-moving average* (ARIMA) process of orders p, d and q, or ARIMA(p,d,q), and x_t is said to be integrated of order d, denoted $I(d)$.

> Several points concerning the ARIMA class of models are of importance. Consider again above formation, with $\theta_0 = 0$ for simplicity:
>
> $$\phi(B)\nabla^d x_t = \theta(B)a_t$$
>
> This process can equivalently be defined by the two equations:
>
> $$\phi(B)w_t = \theta(B)a_t, \quad w_t = \nabla^d x_t$$
>
> so that, as previously noted, the model corresponds to assuming that $\nabla^d x_t$ can be represented by a stationary and invertible ARMA process. Alternatively, for $d \geq 1$, (4.10) can be inverted to give:
>
> $$x_t = S^d w_t$$
>
> where S is the infinite summation, or integral, operator defined by
>
> $$S = \left(1 + B + B^2 + \cdots\right) = (1 - B)^{-1} = \nabla^{-1}$$

This type of nonstationary behavior is often referred to as **Homogenous nonstationarity**: local behavior appears to be roughly independent of level.

If we want to use ARMA models for which the behavior of the process is indeed independent of its level, then the autoregressive polynomial $\phi(B)$ must be chosen so that:

$$\phi(B)\left(x_t + c\right) = \phi(B)x_t$$

where c is any constant. Thus:

$$\phi(B)c = 0$$

Chapter 5: ARIMA Models for Nonstationary Time Series

5.1 Determining The Order of Integration of a Time Series

As we have shown in Chap 2, the order of integration, d, is a crucial determinant of the properties exhibited by a time series. If we restrict ourselves to the most common values of zero and one for d, so that x_t is either $I(0)$ or $I(1)$, then it is useful to bring together the properties of these two processes.

If x_t is $I(0)$, which we will sometimes denote $x_t \sim I(0)$ even though such a notation has been used previously to denote the distributional characteristics of a series, then, if we assume for convenience that x_t has zero mean;

- the variance of x_t is finite and does not depend on t;
- the innovation at has only a temporary effect on the value of x_t;
- the expected length of time between crossings of $x = 0$ is finite, so that x_t fluctuates around its mean of zero;
- the autocorrelations, ρ_k, decrease steadily in magnitude for large enough k, so that their sum is finite.

If, on the other hand, $x_t \sim I(1)$ with %$x_0 = 0$, then;

- the variance of xt goes to infinity as t goes to infinity;
- an innovation at has a permanent effect on the value of xt because xt is the sum of all previous innovations.$x_t = \nabla^{-1} a_t = S a_t = \sum_{i=0}^{t-1} a_{t-i}$;
- the expected time between crossings of $x = 0$ is infinite;
- the autocorrelations $\rho_k \rightarrow 1$ for all k as t goes to infinity.

5.2 Test for a Unit Root

Given the importance of choosing the correct order of differencing, we should have available a formal testing procedure to determine d. To introduce the issues involved in developing such a procedure, we begin by considering the simplest case, that of the zero mean AR(1) process:

$$x_t = \phi x_{t-1} + a_t \quad t = 1, 2, \ldots, T$$

where $a_t \sim \mathrm{WN}\left(0, \sigma^2\right)$ and $x_0 = 0$. The OLS estimator of ϕ is given by

$$\hat{\phi}_T = \frac{\sum_{t=1}^T x_{t-1} x_t}{\sum_{t=1}^T x_t^2}$$

As we have seen, x_t will be $I(0)$ if $|\phi| < 1$, but will be $I(1)$ if $\phi = 1$, so that testing this null hypothesis, that of a “unit root,” becomes a method of determining the correct order of differencing/integration. Given the estimate $\hat{\phi}_T$, a conventional way of testing the null hypothesis would be to construct the t-statistic

$$t_\phi = \frac{\hat{\phi}_T - 1}{\hat{\sigma}_{\hat{\phi}_T}} = \frac{\hat{\phi}_T - 1}{\left(s_T^2 / \sum_{t=1}^T x_{t-1}^2\right)^{1/2}}$$

where

$$\sigma_{\hat{\phi}_T} = \left(\frac{s_T^2}{\sum_{t=1}^T x_{t-1}^2}\right)^{1/2}$$

is the usual OLS standard error for $\hat{\phi}_T$ and s_T^2 is the OLS estimator of σ^2:

$$s_T^2 = \frac{\sum_{t=1}^T \left(x_t - \hat{\phi}_T x_{t-1}\right)^2}{T - 1}$$

5.3 Trend and Difference Stationarity

> 样本时间序列展现了随机变量的历史和现状，因此所谓随机变量基本性态的维持不变也就是要求样本数据时间序列的本质特征仍能延续到未来。我们用样本时间序列的均值、方差、协（自）方差来刻画该样本时间序列的本质特征。于是，我们称这些统计量的取值在未来仍能保持不变的样本时间序列具有平稳性。可见，一个平稳的时间序列指的是：遥想未来所能获得的样本时间序列，我们能断定其均值、方差、协方差必定与眼下已获得的样本时间序列等同。
>
> 相反，如果样本时间序列的本质特征只存在于所发生的当期，并不会延续到未来，亦即样本时间序列的均值、方差、协方差非常数，则这样一个过于独特的时间序列不足以昭示未来，我们便称这样的样本时间序列是非平稳的。
>
> 形象地理解，平稳性就是要求经由样本时间序列所得到的拟合曲线在未来的一段期间内仍能顺着现有的形态“惯性”地延续下去；如果数据非平稳，则说明样本拟合曲线的形态不具有“惯性”延续的特点，也就是基于未来将要获得的样本时间序列所拟合出来的曲线将迥异于当前的样本拟合曲线。

> 可见，时间序列平稳是经典回归分析赖以实施的基本假设；只有基于平稳时间序列的预测才是有效的。如果数据非平稳，则作为大样本下统计推断基础的“一致性”要求便被破坏，基于非平稳时间序列的预测也就失效。

5.4 Estimating Trends Robustly

Consider again the linear trend model: $x_t = \beta_0 + \beta_1 t + \varepsilon_t$, to develop robust tests of trend, we start with the simplest case in which $\varepsilon_t = \rho\varepsilon_{t-1} + a_t$, where $\varepsilon_t \text{ is } I(0) \text{ if } |\rho| < 1 \text{ and } I(1) \text{ if } \rho = 1$. We then wish to test $H_0 : \beta_1 = \beta_1^0$ against the alternative $H_1 : \beta_1 \neq \beta_1^0$. If ε_t is known to be $I(0)$ then an optimal test of H_0 against H_1 is given by the “slope” t-ratio

$$z_0 = \frac{\hat{\beta}_1 - \beta_1^0}{s_0} \quad s_0 = \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{t=1}^T (t - \bar{t})^2}}$$

where $\hat{\sigma}_\varepsilon^2 = (T-2)^{-1} \sum_{t=1}^T \left(x_t - \hat{\beta}_0 - \hat{\beta}_1 t\right)^2$ is the error variance from OLS estimation of x_t.

Under H_0, z_0 will be asymptotically standard normal.

If ε_t is known to be $I(1)$ then the optimal test of H_0 against H_1 is based on the t-ratio associated with the OLS estimator of β_1 in the first-differenced form of x_t,

$$\nabla x_t = \beta_1 + \nu_t \quad t = 2, \ldots, T$$

where $\nu_t = \nabla\varepsilon_t$:

$$z_1 = \frac{\tilde{\beta}_1 - \beta_1^0}{s_1} \quad s_1 = \sqrt{\frac{\tilde{\sigma}_\nu^2}{T-1}}$$

Here

$$\tilde{\beta}_1 = (T-1) \sum_{t=2}^T \nabla x_t = (T-1)\left(x_T - x_1\right)$$

is the OLS estimator of β_1 in ∇x_t and $\tilde{\sigma}_\nu^2 = (T-2)^{-1} \sum_{t=2}^T \left(\nabla x_t - \tilde{\beta}_1\right)^2$.

Again, under H_0, z_0 will be asymptotically standard normal.

Chapter 6: Breaking and Nonlinear Trends

6.1 Breaking Trend Model

Assume, for simplicity, that there is a single break at a known point in time $T_b^c \left(1 < T_b^c < T\right)$, with the superscript “c” denoting the “correct” break date, a distinction that will become important in due course.

The simplest breaking trend model is the “level shift” in which the level of x_t shifts from $\mu_0 \text{ to } \mu_1 = \mu_0 + \mu$ at T_b^c . This may be parameterized as

$$x_t = \mu_0 + (\mu_1 - \mu_0)\mathrm{DU}_t^c + \beta_0 t + \varepsilon_t = \mu_0 + \mu\mathrm{DU}_t^c + \beta_0 t + \varepsilon_t$$

$\text{where } \mathrm{DU}_t^c = 0 \text{ if } t \leq T_b^c \text{ and } \mathrm{DU}_t^c = 1 \text{ if } t > T_b^c$. This shift variable may be written more concisely as

$\mathrm{DU}_t^c = \mathbf{1}\left(t > T_b^c\right)$, where $\mathbf{1}(\cdot)$ is the indicator function, so that it takes the value 1 if the argument is true and 0 otherwise. Another possibility is the "changing growth" model in which the slope of the trend changes from $\beta_0 \text{ to } \beta_1 = \beta_0 + \beta \text{ at } T_b^c$ without a change in level.

In this case, the trend function is joined at the time of the break and is often referred to as a **segmented trend**. This model may be parameterized as

$$x_t = \mu_0 + \beta_0 t + (\beta_1 - \beta_0)\mathrm{DT}_t^c + \varepsilon_t = \mu_0 + \beta_0 t + \beta \mathrm{DT}_t^c + \varepsilon_t$$

where $\mathrm{DT}_t^c = \mathbf{1}\left(t > T_t^c\right)\left(t - T_t^c\right)$ models the shift in growth.

Combined model:

$$\begin{aligned} x_t &= \mu_0 + (\mu_1 - \mu_0)\mathrm{DU}_t^c + \beta_0 t + (\beta_1 - \beta_0)\mathrm{DT}_t^c + \varepsilon_t \\ &= \mu_0 + \mu \mathrm{DU}_t^c + \beta_0 t + \beta \mathrm{DT}_t^c + \varepsilon_t \end{aligned}$$

so that x_t undergoes both a shift in level and slope at T_b^c.

The three models will be **breaking trend-stationary models** in ARMA process($\phi(B)\varepsilon_t = \theta(B)a_t$) .

$$\begin{gathered} \nabla x_t = \beta_0 + \mu \nabla \mathrm{DU}_t^c + \varepsilon_t^* = \beta_0 + \mu \mathrm{D}(\mathrm{TB}^c)_t + \varepsilon_t^* \\ \nabla x_t = \beta_0 + \beta \nabla \mathrm{DT}_t^c + \varepsilon_t^* = \beta_0 + \beta \mathrm{DU}_t^c + \varepsilon_t^* \\ \nabla x_t = \beta_0 + \mu \mathrm{D}(\mathrm{TB}^c)_t + \beta \mathrm{DU}_t^c + \varepsilon_t^* \end{gathered}$$

6.2 Breaking Trends and Unit Root Tests

One way to incorporate such a gradual change into the trend function is to suppose that x_t responds to a trend shock in the same way as it reacts to any other shock. Recalling the ARMA specification for ε_t viz., $\phi(B)\varepsilon_t = \theta(B)a_t$, this would imply that $\psi(B) = \phi(B)^{-1}\theta(B)$, which would be analogous to an "innovation outlier" (IO) model. With this specification, tests for the presence of a unit root can be performed using a direct extension of the ADF regression framework to incorporate dummy variables as appropriate:

$$x_t = \mu^A + \theta^A \mathrm{DU}_t^c + \beta^A t + d^A \mathrm{D}(\mathrm{TB}^c)_t + \phi^A x_{t-1} + \sum_{i=1}^{k} \delta_i \nabla x_{t-i} + a_t$$

$$x_t = \mu^B + \theta^B \mathrm{DU}_t^c + \beta^B t + \gamma^B \mathrm{DT}_t^c + \phi^B x_{t-1} + \sum_{i=1}^{k} \delta_i \nabla x_{t-i} + a_t$$

$$x_t = \mu^C + \theta^C \mathrm{DU}_t^c + \beta^C t + \gamma^C \mathrm{DT}_t^c + d^C \mathrm{D}(\mathrm{TB}^c)_t + \phi^C x_{t-1} + \sum_{i=1}^{k} \delta_i \nabla x_{t-i} + a_t$$

The null hypothesis of a unit root imposes the following parameter restrictions on each model:

$$\begin{gathered} \text{Model (A)} : \phi^A = 1, \theta^A = \beta^A = 0 \\ \text{Model (B)} : \phi^B = 1, \beta^B = \gamma^B = 0 \\ \text{Model (C)} : \phi^C = 1, \beta^C = \gamma^C = 0 \end{gathered}$$

> **UNIT ROOTS TESTS WHEN THE BREAK DATE IS UNKNOWN**

> The procedure set out above is only valid when the break date is known independently of the data, for if a systematic search for a break is carried out then the limiting distributions of the tests are no longer appropriate. Problems also occur if an incorrect break date is selected exogenously, with the tests then suffering size distortions and loss of power.
>
> Consequently, several approaches have been developed that treat the occurrence of the break date as unknown and needing to be estimated. The core is choosing $\hat{T}_b$.
>
> Two data-dependent methods for choosing T^ b have been considered, both of which involve estimating the appropriate detrended AO regression or IO regression, for all possible break dates.
>
> - The first method chooses $\hat{T}_b$ as the break date that is most likely to reject the unit root hypothesis, which is the date for which the t-statistic for testing $\phi = 1$ is minimized (i.e., is most negative).
> - The second approach involves choosing $\hat{T}_b$ as the break date for which some statistic that tests the significance of the break parameters is maximized. This is equivalent to minimizing the residual sum of squares across all possible regressions, albeit after some preliminary trimming has been performed, that is, if only break fractions $\tau = T_b/T$ between $0 < \tau_{\min}, \tau_{\max} < 1$ are considered.
>
> Having selected $\hat{T}_b$ by one of these methods, the procedure may then be applied conditional upon this choice.

6.3 Robust Tests for a Breaking Trend

If the break date is known to be at Tc b with break fraction τc then, focusing on the segmented trend model (B),

$$t_\lambda = \lambda\left(S_0\left(\tau^c\right), S_1\left(\tau^c\right)\right) \times \left|t_0\left(\tau^c\right)\right| + \left(1 - \lambda\left(S_0\left(\tau^c\right), S_1\left(\tau^c\right)\right)\right) \times \left|t_1\left(\tau^c\right)\right|$$

with the weight function now being defined as

$$\lambda\left(S_0\left(\tau^c\right), S_1\left(\tau^c\right)\right) = \exp\left(-\left(500 S_0\left(\tau^c\right) S_1\left(\tau^c\right)\right)^2\right)$$

Here $S_0\left(\tau^c\right)$ and $S_1\left(\tau^c\right)$ are KPSS statistics computed from the residuals of

$$x_t = \mu_0 + (\mu_1 - \mu_0)\mathrm{DU}_t^c + \beta_0 t + \varepsilon_t = \mu_0 + \mu\mathrm{DU}_t^c + \beta_0 t + \varepsilon_t$$

and

$$\nabla x_t = \beta_0 + \beta \nabla \mathrm{DT}_t^c + \varepsilon_t^* = \beta_0 + \beta \mathrm{DU}_t^c + \varepsilon_t^*$$

Under $H_0 : \beta = 0, t_\lambda$ will be asymptotically standard normal.

6.4 Confidence Intervals for The Break Date and Multiple Breaks

When the break date is estimated it is often useful to be able to provide a confidence interval for the unknown T_b^c . Perron and Zhu (2005) show that for the segmented trend model (B) and $I(1)$ errors

$$\sqrt{T}\left(\hat{\tau} - \tau^c\right) \overset{d}{\sim} N\left(0, 2\sigma^2/15\beta^2\right)$$

while for $I(0)$ errors

$$T^{3/2}\left(\tilde{\tau}-\tau^{c}\right) \overset{d}{\sim} N\left(0, 4\sigma^{2} / \left(\tau^{c}\left(1-\tau^{c}\right)\beta^{2}\right)\right)$$

so that, for example, a 95% confidence interval for τ^c when the errors are $I(1)$ is given by

$$\hat{\tau} \pm 1.96\sqrt{\frac{2\hat{\sigma}^{2}}{15T\hat{\beta}^{2}}}$$

6.5 Nonlinear Trends

Logistic smooth transition (LSTR):

$$S_t(\gamma, m)=(1+\exp(-\gamma(t-mT)))^{-1}$$

Exponential smooth transition (ESTR):

$$S_t(\gamma, m)=1-\exp\left(-\gamma(t-mT)^{2}\right)$$

Analogous to model A, B, C, three alternative smooth transition trend models may then be specified as

$$\begin{aligned}
x_t &= \mu_0+\mu S_t(\gamma, m)+\varepsilon_t \\
x_t &= \mu_0+\beta_0 t+\mu S_t(\gamma, m)+\varepsilon_t \\
x_t &= \mu_0+\beta_0 t+\mu S_t(\gamma, m)+\beta t S_t(\gamma, m)+\varepsilon_t
\end{aligned}$$

Chapter 7: An Introduction to Forecasting With Univariate Models

7.1 Forecasting With ARIMA Models

An important feature of the univariate models is their ability to provide forecasts of future values of the observed series.

There are two aspects to forecasting: the provision of a forecast for a future value of the series and the provision of a forecast error that can be attached to this point forecast. This forecast error may then be used to construct forecast intervals to provide an indication of the precision these forecasts are likely to possess.

Process:

- To formalize the forecasting problem, suppose we have a realization $\left(x_{1-d}, x_{2-d}, \cdots, x_T\right)$ from a general ARIMA (p,d,q) process

$$\phi(B)\nabla^{d} x_t=\theta_0+\theta(B) a_t$$

- and that we wish to forecast a future value x_T+h, h being known as the lead time or forecast horizon. If we let

$$\alpha(B)=\phi(B)\nabla^{d}=\left(1-\alpha_1 B-\alpha_2 B^{2}-\cdots-\alpha_{p+d} B^{p+d}\right)$$

  then Eq.1 becomes, for time $T+h$, $\alpha(B) x_{T+h}=\theta_0+\theta(B) a_{T+h}$, that is,

$$\begin{aligned}
x_{T+h} =& \alpha_1 x_{T+h-1}+\alpha_2 x_{T+h-2}+\cdots+\alpha_{p+d} x_{T+h+p-d}+\theta_0+a_{T+h} \\
& -\theta_1 a_{T+h-1}-\cdots-\theta_q a_{T+h-q}
\end{aligned}$$

- Clearly, observations from $T+1$ onwards are unavailable, but a **minimum mean square error (MMSE)** forecast of x_{T+h} made at time T (known as the origin), and denoted $f_{T,h}$, is given by the conditional expectation

$$\begin{aligned} f_{T,h} = & E\left(\alpha_1 x_{T+h-1} + \alpha_2 x_{T+h-2} + \cdots + \alpha_{p+d} x_{T+h-p-d} + \theta_0 \right. \\ & \left. + a_{T+h} - \theta_1 a_{T+h-1} - \cdots - \theta_q a_{T+h-q} \mid x_T, x_{T-1}, \ldots\right). \end{aligned}$$

- This is the forecast that will minimize the variance of the h-step ahead forecast error $e_{T,h} = x_{T+h} - f_{T,h}$ that is, it will minimize $E\left(e_{T,h}^2\right)$. Now it is clear that

$$\begin{aligned} E\left(x_{T+j} \mid x_T, x_{T-1}, \ldots\right) &= \begin{cases} x_{T+j}, & j \leq 0 \\ f_{T,j}, & j > 0\text{’} \end{cases} \\ E\left(a_{T+j} \mid x_T, x_{T-1}, \ldots\right) &= \begin{cases} a_{T+j}, & j \leq 0 \\ 0, & j > 0\text{’} \end{cases} \end{aligned}$$

so that, to evaluate $f_{T,h}$, all we need to do is:

1. Replace past expectations $(j \leq 0)$ by known values, x_{T+j} and a_{T+j}
2. Replace future expectations $(j > 0)$ by forecast values, $f_{T,j}$ and 0

forecast interval may be constructed as $f_{T,h} \pm z_{\varsigma/2}\sqrt{V\left(e_{T,h}\right)}$, where $z_{\varsigma/2}$ is the $\varsigma/2$ percentage point of the standard normal distribution.

7.2 Forecasting a Trend Stationary Process

Consider the trend stationary (TS) process

$$x_t = \beta_0 + \beta_1 t + \varepsilon_t \quad \phi(B)\varepsilon_t = \theta(B) a_t$$

The forecast of x_{T+h} made at time T is

$$f_{T,h} = \beta_0 + \beta_1(T+h) + g_{T,h}$$

where $g_{T,h}$ is the forecast of ε_{T+h}, which from $f_{T,h}$ is given by:

$$\begin{aligned} g_{T,h} = & E\left(\phi_1 \varepsilon_{T+h-1} + \phi_2 \varepsilon_{T+h-2} + \cdots + \phi_p x_{T+h-p} + a_{T+h} \right. \\ & \left. -\theta_1 a_{T+h-1} - \cdots - \theta_q a_{T+h-q} \mid \varepsilon_T, \varepsilon_{T-1}, \ldots\right) \end{aligned}$$

Since ε_t is stationary, we know that $g_{T,h} \rightarrow 0 \text{ as } h \rightarrow \infty$. Thus, for large h, $f_{T,h} = \beta_0 + \beta_1(T+h)$ and forecasts will be given simply by the extrapolated linear trend. For smaller h there will also be the component $g_{T,h}$, but this will dissipate as h increases. The forecast error will be

$$e_{T,h} = x_t - f_{T,h} = \varepsilon_{T+h} - g_{T,h}$$

and, hence, the uncertainty in any TS forecast is due solely to the error in forecasting the ARMA component.

Chapter 8: Unobserved Component Models, Signal Extraction, and Filters

8.1 Unobserved Component Models

A difference stationary, that is, $I(1)$, time series may always be decomposed into a stochastic nonstationary trend, or signal, component and a stationary noise, or irregular, component:

$$x_t = z_t + u_t$$

Such a decomposition can be performed in several ways. For instance, Muth's (1960) classic example assumes that the trend component z_t is a random walk

$$z_t = \mu + z_{t-1} + v_t$$

while ut is white noise and independent of $\mathcal{V}_t$, that is, $u_t \sim \mathbf{WN}\left(0, \sigma_u^2\right)$ and $v_t \sim \mathrm{WN}\left(0, \sigma_v^2\right)$, with $E\left(u_t v_{t-i}\right) = 0$ for all i. Thus, it follows that ∇x_t is the stationary process

$$\nabla x_t = \mu + v_t + u_t - u_{t-1}$$

Models of the form above are known as **unobserved component (UC)** models. a more general formulation for the components being:

$$\begin{aligned} \nabla z_t &= \mu + \gamma(B) v_t \\ u_t &= \lambda(B) a_t \end{aligned}$$

where v_t and a_t are independent white noise sequences with finite variances σ_v^2 v and σ_a^2, and where $\gamma(B)$ and $\lambda(B)$ are stationary polynomials having no common roots. It can be shown that x_t will then have the form:

$$\nabla x_t = \mu + \theta(B) e_t$$

where $\theta(B)$ and σ_e^2 can be obtained from:

$$\sigma_e^2 \frac{\theta(B)\theta\left(B^{-1}\right)}{(1-B)\left(1-B^{-1}\right)} = \sigma_v^2 \frac{\gamma(B)\gamma\left(B^{-1}\right)}{(1-B)\left(1-B^{-1}\right)} + \sigma_a^2 \lambda(B)\lambda\left(B^{-1}\right)$$

8.2 Signal Extraction

Given a UC model of the form of $x_t = z_t + u_t$ and models for z_t and u_t, it is often useful to provide estimates of these two unobserved components, a procedure that is known as **signal extraction**.

MMSE estimate of z_t:

$$\begin{gathered} \hat{z}_t = \nu_z(B) x_t = \sum_{j=-\infty}^{\infty} \nu_{zj} x_{t-j} \\ \hat{u}_t = x_t - \hat{z}_t = (1 - \nu_z(B)) x_t = \nu_u(B) x_t \end{gathered}$$

where the filter $\nu_z(B)$ is defined as:

$$\nu_z(B) = \frac{\sigma_v^2 \gamma(B) \gamma\left(B^{-1}\right)}{\sigma_e^2 \theta(B) \theta\left(B^{-1}\right)}$$

8.3 Filters: Low-pass, High-pass, Band-pass

Hodrick-Prescott trend filter: This filter is derived by minimizing the variation in the noise component $u_t = x_t - z_t$, subject to a condition on the "smoothness" of the trend component z_t.

This smoothness condition penalizes acceleration in the trend, so that the minimization problem becomes that of minimizing the function:

$$\sum_{t=1}^{T} u_t^2 + \lambda \sum_{t=1}^{T} \left(\left(z_{t+1} - z_t\right) - \left(z_t - z_{t-1}\right)\right)^2$$

with respect to $z_t, t = 0, 1, \ldots, T+1$, where λ is a Lagrangean multiplier that can be interpreted as a smoothness parameter. The higher the value of λ, the smoother the trend is, so that in the limit, as $\lambda \to \infty$, z_t becomes a linear trend. The first-order conditions are:

$$\begin{aligned} 0 = & -2\left(x_t - z_t\right) + 2\lambda\left(\left(z_t - z_{t-1}\right) - \left(z_{t-1} - z_{t-2}\right)\right) - 4\lambda\left(\left(z_{t+1} - z_t\right) - \left(z_t - z_{t-1}\right)\right) \\ & + 2\lambda\left(\left(z_{t+2} - z_{t+1}\right) - \left(z_{t+1} - z_t\right)\right) \end{aligned}$$

which may be written as:

$$x_t = z_t + \lambda(1-B)^2\left(z_t - 2z_{t+1} + z_{t+2}\right) = \left(1 + \lambda(1-B)^2\left(1 - B^{-1}\right)^2\right) z_t$$

so that the Hodrick-Prescott (HP) trend estimator is

$$\hat{z}_t(\lambda) = \left(1 + \lambda(1-B)^2\left(1 - B^{-1}\right)^2\right)^{-1} x_t$$

The MMSE trend estimator can be written as:

$$\hat{z}_t = \frac{\sigma_\nu^2 \gamma(B) \gamma\left(B^{-1}\right)}{\sigma_e^2 \theta(B) \theta\left(B^{-1}\right)} x_t = \frac{\gamma(B) \gamma\left(B^{-1}\right)}{\gamma(B) \gamma\left(B^{-1}\right) + \left(\sigma_a^2 / \sigma_\nu^2\right) \lambda(B) \lambda\left(B^{-1}\right)} x_t$$

In filtering terminology the HP filter (8.18) is a *low-pass filter*.

In general, $a(\omega) = |a(\omega)| e^{-i\theta(\omega)}$, where:

$$\theta(\omega) = \tan^{-1} \frac{\sum_j a_j \sin \omega j}{\sum_j a_j \cos \omega j}$$

is the **phase shift**, indicating the extent to which the ωfrequency component of x_t is displaced in time.

Frequency response function:

$$a(\omega) = \sum_i e^{-i\omega j}, \quad 0 \leq \omega \leq 2\pi$$

Power transfer function:

$$|a(\omega)|^2 = \left(\sum_j a_j \cos \omega j\right)^2 + \left(\sum_j a_j \sin \omega j\right)^2$$

the **gain** is defined as $|a(\omega)|$, measuring the extent to which the amplitude of the ω-frequency component of x_t is altered through the filtering operation.

Chapter 9: Seasonality and Exponential Smoothing

9.1 Modeling Deterministic Seasonality

Seasonal mean model: different mean for each season

$$x_t = \sum_{i=1}^{m} \alpha_i s_{i,t} + \varepsilon_t$$

where the seasonal dummy variable $s_{i,t}$ takes the value 1 for the ith season and zero otherwise, there being m seasons in the year.

9.2 Modeling Stochastic Seasonality

stochastic seasonality: ARIMA processes

An important consideration when attempting to model a seasonal time series with an ARIMA model is to determine what sort of process will best match the SACFs and PACFs that characterize the data.

9.3 Mixed Seasonal Models

The deterministic and stochastic seasonal models, may be combined to form, on setting $d = D = 1$ for both simplicity and because these are the settings that are typically found,

$$x_t = \sum_{i=1}^{m} \alpha_i s_{i,t} + \frac{\theta_q(B)\Theta_Q\left(B^m\right)}{\phi_p(B)\Phi_P\left(B^m\right)\nabla\nabla_m} a_t$$

where $\alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$.

9.4 Seasonal Adjustment

Remove the seasonal component, rather than modeling it as an integral part of the stochastic process generating the data, as in fitting a seasonal ARIMA model,

9.5 Exponential Smoothing

For two-component UC model, where $x_t = z_t + u_t$, then a simple model for the signal or “level” z_t is to assume that its current value is an exponentially weighted moving average of current and past observations of x_t:

$$\begin{aligned} z_t &= \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \cdots = \alpha \sum_{j=0}^{\infty}(1-\alpha)^j x_{t-j} \\ &= \alpha\left(1 + (1-\alpha)B + (1-\alpha)^2B^2 + \cdots + (1-\alpha)^jB^j + \cdots\right)x_t \end{aligned}$$

or

$$\begin{aligned} (1-(1-\alpha)B)z_t &= \alpha x_t \\ z_t &= \alpha x_t + (1-\alpha)z_{t-1} \end{aligned}$$

Chapter 10: Volatility and Generalized Autoregressive Conditional Heteroskedastic Processes

10.1 Volatility

A stochastic model having time-varying conditional variances may be defined by supposing that xt is generated by the **product process**:

$$x_t = \mu + \sigma_t U_t$$

where U_t is a standardized process, so that $E\left(U_t\right) = 0$ and $V\left(U_t\right) = E\left(U_t^2\right) = 1$ for all t, and σ_t is a sequence of positive random variables such that:

$$V\left(x_t \mid \sigma_t\right)=E\left(\left(x_t-\mu\right)^2 \mid \sigma_t\right)=\sigma_t^2 E\left(U_t^2\right)=\sigma_t^2$$

σ_t^2 is the conditional variance and σt the conditional standard deviation of x_t.

$$E(x_t-\mu)^2=E\left(\sigma_t^2 U_t^2\right)=E\left(\sigma_t^2\right) E\left(U_t^2\right)=E\left(\sigma_t^2\right)$$

and autocovariances

$$E\left(x_t-\mu\right)\left(x_{t-k}-\mu\right)=E\left(\sigma_t \sigma_{t-k} U_t U_{t-k}\right)=E\left(\sigma_t \sigma_{t-k}\right) E\left(U_t U_{t-k}\right)=0$$

10.2 Autoregressive Conditional Heteroskedastic Process

First-order autoregressive conditional heteroskedastic [ARCH(1)] process:

Consider σ_t^2 where they are a function of past values of x_t:

$$\sigma_t^2=f\left(x_{t-1}, x_{t-2}, \ldots\right)$$

A simple example is:

$$\sigma_t^2=f\left(x_{t-1}\right)=\alpha_0+\alpha_1(x_{t-1}-\mu)^2$$

where α_0 and α_1 are both positive to ensure that $\sigma_t^2>0$ With $U_t \sim \mathrm{NID}(0,1)$ and independent of $\sigma_t, x_t=\mu+\sigma_t U_t$ is then conditionally normal:

$$x_t \mid x_{t-1}, x_{t-2}, \ldots \sim \mathrm{N}\left(\mu, \sigma_t^2\right)$$

so that

$$V\left(x_t \mid x_{t-1}\right)=\alpha_0+\alpha_1(x_{t-1}-\mu)^2$$

A more convenient notation is to define $\varepsilon_t=x_t-\mu=U_t \sigma_t$, so that the ARCH(1) model can be written as:

$$\begin{gathered}\varepsilon_t \mid x_{t-1}, x_{t-2}, \ldots \sim \mathrm{NID}\left(0, \sigma_t^2\right) \\ \sigma_t^2=\alpha_0+\alpha_1 \varepsilon_{t-1}^2\end{gathered}$$

On defining $\nu_t=\varepsilon_t^2-\sigma_t^2$, the model can also be written as:

$$\varepsilon_t^2=\alpha_0+\alpha_1 \varepsilon_{t-1}^2+\nu_t$$

Since $E\left(\nu_t \mid x_{t-1}, x_{t-2}, \ldots\right)=0$, the model corresponds directly to an AR(1) model for the squared innovations ε_t^2.

A natural extension is to the ARCH(q) process:

$$\sigma_t^2=f\left(x_{t-1}, x_{t-2}, \ldots, x_{t-q}\right)=\alpha_0+\sum_{i=1}^q \alpha_i(x_{t-i}-\mu)^2$$

generalized ARCH (GARCH) process:

$$\begin{aligned}\sigma_t^2 & =\alpha_0+\sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2+\sum_{i=1}^p \beta_i \sigma_{t-i}^2 \\ & =\alpha_0+\alpha(B) \varepsilon_{t-1}^2+\beta(B) \sigma_{t-1}^2\end{aligned}$$

where $p>0 \quad \text{and} \quad \beta_i \geq 0, \quad i \leq 1 \leq p.$

For the GARCH (1,1) process,

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

The equivalent form of the GARCH(p,q) process is:

$$\varepsilon_t^2 = \alpha_0 + (\alpha(B) + \beta(B))\varepsilon_{t-1}^2 + \nu_t - \beta(B)\nu_{t-1}$$

so that ε_t^2 is ARMA(m,p), where $m = \max(p, q)$. *This process will be weakly stationary if the roots of* $1 - \alpha(B) - \beta(B)$ *are all less than unity, so that* $\alpha(1) + \beta(1) < 1$.

10.3 Testing for The Presence of Arch Errors

Suppose that an ARMA model for x_t has been estimated, from which the residuals e_t have been obtained.

If ε_t is GARCH (p,q) then ε_t^2 is ARMA(m,p), where $m = \max(p, q)$.squared residuals e2 t from the estimation of a pure ARMA process can then be used to identify m and p, and therefore q, in a similar fashion to the way the residuals themselves are used in conventional ARMA modeling.

Formal tests are also available. A test of the null hypothesis that εt has a constant conditional variance against the alternative that the conditional variance is given by an ARCH(q) process, which is a test of $\alpha_1 = \ldots = \alpha_q = 0$ conditional upon $\beta_1 = \ldots = \beta_p = 0$, may be based on the Lagrange Multiplier (LM) principle. The test procedure is to run a regression of $e_t^2 \text{ on } e_{t-1}^2, \ldots, e_{t-q}^2$ and to test the statistic $T \cdot R^2$ as a χ_q^2 variate, where R^2 is the squared multiple correlation coefficient of the regression.

Exponential GARCH (EGARCH) model:

$$\log\left(\sigma_t^2\right) = \alpha_0 + \alpha_1 g\left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right) + \beta_1 \log\left(\sigma_{t-1}^2\right)$$

where

$$g\left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right) = \theta_1 \frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \left(\left|\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right| - E\left|\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right|\right)$$

Nonlinear ARCH model:

$$\sigma_t^\gamma = \alpha_0 + \alpha_1 g^\gamma\left(\varepsilon_{t-1}\right) + \beta_1 \sigma_{t-1}^\gamma$$

while an alternative is the threshold ARCH process:

$$\sigma_t^\gamma = \alpha_0 + \alpha_1 h^{(\gamma)}\left(\varepsilon_{t-1}\right) + \beta_1 \sigma_{t-1}^\gamma$$

where

$$h^{(\gamma)}\left(\varepsilon_{t-1}\right) = \theta_1 |\varepsilon_{t-1}|^\gamma \mathbf{1}\left(\varepsilon_{t-1} > 0\right) + |\varepsilon_{t-1}|^\gamma \mathbf{1}\left(\varepsilon_{t-1} \leq 0\right)$$

10.4 Forecasting From an ARMA-GARCH Model

Suppose we have the ARMA(P,Q)-GARCH(p,q) model

$$\begin{gathered} x_t = \Phi_1 x_{t-1} + \cdots + \Phi_P x_{t-P} + \Theta_0 + \varepsilon_t - \Theta_1 \varepsilon_{t-1} - \cdots - \Theta_Q \varepsilon_{t-Q} \\ \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2 \end{gathered}$$

Chapter 11: Nonlinear Stochastic Processes

11.1 Martingales, Random Walks, and Nonlinearity

Martingale: A martingale may be defined as a stochastic process x_t having the following properties:

- $E\left(\left|x_t\right|\right)<\infty$ for each t
- $E\left(x_t \mid x_s, x_{s-1}, \ldots\right)=x_s$

written as

$$E\left(x_t-x_s \mid x_s, x_{s-1}, \ldots\right)=0, \quad s<t$$

the martingale property implies that the MMSE forecast of a future increment of a martingale is zero. This property can be generalized to situations where:

$$E\left(x_t-x_s \mid x_s, x_{s-1}, \ldots\right) \geq 0, \quad s<t$$

in which we have a **sub-martingale,** and to the case where this inequality is reversed, giving us a **super-martingale**.

The martingale can be written equivalently as:

$$x_t=x_{t-1}+a_t$$

where a_t is known as the martingale increment or **martingale difference**.

11.2 Nonlinear Stochastic Models

A stochastic process can then be considered nonlinear if it does not satisfy the assumptions underlying the decomposition.

11.3 Bilinear Models

General form:

$$\phi(B)\left(x_t-\mu\right)=\theta(B) \varepsilon_t+\sum_{i=1}^R \sum_{i=1}^S \gamma_{i j} x_{t-i} \varepsilon_{t-j}$$

Here $\varepsilon_t \sim SWN\left(0, \sigma_{\varepsilon}^2\right)$, where this notation is used to denote that the innovations ε_t are strict white noise. The second term on the right hand side is a bilinear form in ε_{t-j} and x_{t-i}, and this accounts for the nonlinear character of the model, for if all the $\gamma_{i j}$ are zero, the equation clearly reduces to the familiar ARMA model.

Diagonal model:

$$x_t=\varepsilon_t+\gamma_{i j} x_{t-i} \varepsilon_{t-j}$$

If $i>j$ the model is called s**uper-diagonal**, if $i=j$ it is **diagonal**, and if $i<j$, it is **sub-diagonal**. If we define $\lambda=\gamma_{i j} \sigma$ then, for super-diagonal models, x_t has zero mean and variance $\sigma_{\varepsilon}^2 /\left(1-\lambda^2\right)$, so that $|\lambda|<1$ is a necessary condition for stability.

11.4 Threshold and Smooth Transition Autoregressions

Self-exciting threshold autoregressive (SETAR) process: allows for asymmetry by defining a set of piecewise autoregressive models whose switch points, or "thresholds," are generally unknown:

$$x_t = \sum_{j=1}^{r} (\phi_{j,1} x_{t-1} + \cdots + \phi_{j,p} x_{t-p} + a_{j,t}) \mathbf{1}\left(c_{j-1} < x_{t-d} \leq c_j\right)$$

Here d is the (integer-valued) delay parameter and $c_1 < c_2 < \cdots < c_{r-1}$ are the thresholds: the model is often denoted SETAR(r: p, d). It is assumed that $a_{j,t} \sim WN\left(0, \sigma_j^2\right), j = 1, \ldots, r$, so that the error variance is allowed to alter across the r "regimes."

exponential autoregressive (EAR) process:

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + G\left(\gamma, x_{t-d}\right)\left(\varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p}\right) + a_t$$

where the transition function

$$G\left(\gamma, x_{t-d}\right) = \exp\left(-\gamma x_{t-d}^2\right), \quad \gamma > 0$$

is symmetric around zero, where it takes the value unity, and as $|x_{t-d}| \to \infty$ so $G\left(\gamma, x_{t-d}\right) \to 0$. The EAR may be interpreted as a linear AR process with stochastic time-varying coefficients $\phi_i + G\left(\gamma, x_{t-d}\right)\varphi_i$.

11.5 Markrov-Switching Models

The setup is that of the UC model, where z_t now evolves as a two-state Markov process:

$$z_t = \alpha_0 + \alpha_1 S_t$$

where

$$\begin{aligned} P\left(S_t = 1 \mid S_{t-1} = 1\right) &= p \\ P\left(S_t = 0 \mid S_{t-1} = 1\right) &= 1 - p \\ P\left(S_t = 1 \mid S_{t-1} = 0\right) &= 1 - q \\ P\left(S_t = 0 \mid S_{t-1} = 0\right) &= q \end{aligned}$$

The noise component u_t is assumed to follow an AR(r) process $\phi(B) u_t = \varepsilon_t$, where the innovation sequence ε_t is strict white noise but $\phi(B)$ may contain a unit root.

The stochastic process for S_t is strictly stationary, having the AR(1) representation:

$$S_t = (1 - q) + \lambda S_{t-1} + V_t$$

where $\lambda = p + q - 1$ and where the innovation V_t has the conditional probability distribution

$$\begin{aligned} P\left(V_t = (1 - p) \mid S_{t-1} = 1\right) &= p \\ P\left(V_t = -p \mid S_{t-1} = 1\right) &= 1 - p \\ P\left(V_t = -(1 - q) \mid S_{t-1} = 0\right) &= q \\ P\left(V_t = q \mid S_{t-1} = 0\right) &= 1 - q \end{aligned}$$

This innovation is uncorrelated with lagged values of S_t, since

$$E\left(V_t \mid S_{t-j} = 1\right) = E\left(V_t \mid S_{t-j} = 0\right) = 0 \quad \text{for } j \geq 1$$

but it is not independent of such lagged values, as, for example,

$$E\left(V_t^2 \mid S_{t-1}=1\right)=p(1-p)$$
$$E\left(V_t^2 \mid S_{t-1}=0\right)=q(1-q)$$

The variance of the Markov process can be shown to be

$$\alpha_1^2 \frac{(1-p)(1-q)}{(2-p-q)^2}$$

As this variance approaches zero, i.e., as p and q approach unity, so the random walk component approaches a deterministic trend. If $\phi(B)$ contains no unit roots, x_t will approach a trend stationary process, whereas if $\phi(B)$ does contain a unit root, x_t approaches a difference stationary process.

11.6 Nonlinear Dynamics and Chaos

Chaos:

> An example of a chaotic process is one that is generated by a deterministic difference equation
>
> $$x_t=f\left(x_{t-1}, \ldots, x_{t-p}\right)$$
>
> such that x_t does not tend to a constant or a (limit) cycle and has estimated covariances that are extremely small or zero. A simple example is provided by Brock (1986), where a formal development of deterministic chaos models is provided. Consider the difference equation
>
> $$x_t=f\left(x_{t-1}\right), \quad x_0 \in[0,1]$$
>
> where
>
> $$f(x)= \begin{cases}x / \alpha & x \in[0, \alpha] \\ (1-x) /(1-\alpha) & x \in[\alpha, 1] \quad 0<\alpha<1\end{cases}$$
>
> Most realizations (or trajectories) of this difference equation generate the same SACFs as an AR(1) process for x_t with parameter $\phi=(2 \alpha-1)$. Hence, for $\alpha=0.5$, the realization will be indistinguishable from white noise, even though it has been generated by a purely deterministic nonlinear process. Hsieh (1991) provides further discussion of this function, called a **tent map** because the graph of x_t against x_{t-1} (known as the phase diagram) is shaped like a “tent.” Hsieh also considers other relevant examples of chaotic systems, such as the **logistic map**
>
> $$x_t=4 x_{t-1}\left(1-x_{t-1}\right)=4 x_{t-1}-4 x_{t-1}^2, \quad 0<x_0<1$$
>
> This also has the same autocorrelation properties as white noise, although x_t^2 has an SACF consistent with an MA(1) process.

11.7 Testing for Nonlinearity

Regression Error Specification Test (RESET): constructed from the auxiliary regression

$$e_t=\sum_{i=1}^p \varphi_i x_{t-i}+\sum_{j=2}^h \delta_j \hat{x}_t^j+v_t$$

and is the F-test of the hypothesis $H_0: \delta_i=0, j=2, \ldots, h.$ If $h=2$, this is equivalent to Keenan’s test, while Tsay augments the auxiliary regression with second-order terms:

$$e_t = \sum_{i=1}^{p} \varphi_i x_{t-i} + \sum_{i=1}^{p} \sum_{j=i}^{p} \delta_{ij} x_{t-i} x_{t-j} + v$$

in which the linearity hypothesis is $H_0 : \delta_{ij} = 0$, for all i and j.

Chapter 12: Transfer Functions and Autoregressive Distributed Lag Modeling

12.1 Transfer Function-noise Models

Single-input transfer function-noise:

$$y_t = v(B)x_t + n_t$$

where the lag polynomial $v(B) = v_0 + v_1 B + v_2 B^2 + \cdots$ allows x to influence y via a **distributed lag**: $v(B)$ is often referred to as the **transfer function** and the coefficients v_i as the **impulse response weights**.

It is assumed that both input and output variables are stationary, perhaps after appropriate transformation.

$v(B)$ may be written as the **rational distributed lag**

$$v(B) = \frac{\omega(B)B^b}{\delta(B)}$$

Here the numerator and denominator polynomials are defined as

$$\begin{gathered} \omega(B) = \omega_0 - \omega_1 B - \cdots - \omega_s B^s \\ \delta(B) = 1 - \delta_1 B - \cdots - \delta_r B^r \end{gathered}$$

with the roots of $\delta(B)$ all assumed to be less than unity.

The model can be written as:

$$y_t = \sum_{j=1}^{M} v_j(B) x_{j,t} + n_t = \sum_{j=1}^{M} \frac{\omega_j(B)B^{b_j}}{\delta_j(B)} x_{j,t} + \frac{\theta(B)}{\phi(B)} a_t$$

where

$$\begin{gathered} \omega_j(B) = \omega_{j,0} - \omega_{j,1} B - \cdots - \omega_{j,s_j} B^{s_j} \\ \delta_j(B) = 1 - \delta_{j,1} B - \cdots - \delta_{j,r_j} B^{r_j} \end{gathered}$$

12.2 Autoregressive Distributed Lag Models

Autoregressive distributed lag (ARDL) model:

$$y_t = \sum_{j=1}^{M} v_j(B) x_{j,t} + n_t = \sum_{j=1}^{M} \frac{\omega_j(B)B^{b_j}}{\delta_j(B)} x_{j,t} + \frac{\theta(B)}{\phi(B)} a_t$$

where

$$\delta_1(B) = \cdots = \delta_M(B) = \phi(B) \quad \theta(B) = 1$$

so that the model is, on defining $\beta_j(B) = \omega_j(B)B^{b_j}$ and including an intercept,

$$\phi(B)y_t = \beta_0 + \sum_{j=1}^{M} \beta_j(B)x_{j,t} + a_t$$

Chapter 13: Vector Autoregressions and Granger Causality

13.1 Multivariate Dynamic Regression Models

Multivariate dynamic regression: ARDL model with two endogenous variables:

$$\begin{aligned} y_{1,t} &= c_1 + a_{11}y_{1,t-1} + a_{12}y_{2,t-1} + b_{10}x_t + b_{11}x_{t-1} + u_{1,t} \\ y_{2,t} &= c_2 + a_{21}y_{1,t-1} + a_{22}y_{2,t-1} + b_{20}x_t + b_{21}x_{t-1} + u_{2,t} \end{aligned}$$

The variances of the two innovations then being

$$E\left(u_1^2\right) = \sigma_1^2 \text{ and } E\left(u_2^2\right) = \sigma_2^2$$

Generalized multivariate dynamic regression: containing n endogenous variables and k exogenous variables.

Gathering these together in the vectors $\mathbf{y}_t' = (y_{1,t}, y_{2,t}, \ldots, y_{n,t}) \text{ and } \mathbf{x}_t' = (x_{1,t}, x_{2,t}, \ldots, x_{k,t})$

The general form of the multivariate dynamic regression model may be written as:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \sum_{i=0}^{q} \mathbf{B}_i \mathbf{x}_{t-i} + \mathbf{u}_t$$

where there is a maximum of p lags on the endogenous variables and a maximum of q lags on the exogenous variables.

Here $\mathbf{c}' = (c_1, c_2, \ldots, c_n)$ is a $1 \times n$ vector of constants and $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_p$ and $\mathbf{B}_0, \mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_q$ are sets of $n \times n$ and $n \times k$ matrices of regression coefficients, respectively, such that

$$\mathbf{A}_i = \begin{bmatrix} a_{11,i} & a_{12,i} & \ldots & a_{1n,i} \\ a_{21,i} & a_{22,i} & \ldots & a_{2n,i} \\ \vdots & & & \vdots \\ a_{n1,i} & a_{n2,i} & \ldots & a_{nn,i} \end{bmatrix} \quad \mathbf{B}_i = \begin{bmatrix} b_{11,i} & b_{12,i} & \ldots & b_{1k,i} \\ b_{21,i} & b_{22,i} & \ldots & b_{2k,i} \\ \vdots & & & \vdots \\ b_{n1,i} & b_{n2,i} & \ldots & b_{nk,i} \end{bmatrix}$$

$\mathbf{u}_t' = (u_{1,t}, u_{2,t}, \ldots, u_{n,t})$ is a $1 \times n$ zero mean vector of innovations (or errors), whose variances and covariances can be gathered together in the symmetric n 3 n error covariance matrix

$$\boldsymbol{\Omega} = E\left(\mathbf{u}_t \mathbf{u}_t'\right) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & & & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$$

It is assumed that these errors are mutually serially uncorrelated, so that $E\left(\mathbf{u}_t \mathbf{u}_s'\right) = \mathbf{0}$ for $t \neq s$, where $\mathbf{0}$ is an $n \times n$ null matrix.

13.2 Vector Autoregressions

Suppose the generalized model does not contain any exogenous variables, so that all the $\mathbf{B}_i$ matrices are zero, and that there are p lags of the endogenous variables in every equation:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{u}_t$$

Because (13.3) is now simply a pth order autoregression in the vector y_t it is known as a **vector autoregression (VAR(p))** of dimension n and, again, can be estimated by multivariate least squares.

where

$$\mathbf{A}(B) = \mathbf{I}_n - \mathbf{A}_1 B - \cdots - \mathbf{A}_p B^p = \mathbf{0}$$

13.3 Granger Causality

Granger causality:

In the VAR the presence of nonzero off-diagonal elements in the $\mathbf{A}_i$ matrices, $a_{rs,i} \neq 0, r \neq s$, implies that there are dynamic relationships between the variables, otherwise the model would collapse to a set of n univariate AR processes.

Granger-cause:

If $a_{rs,i} = 0$ for all $i = 1, 2, \ldots, p$, The variable y_s *does not Granger-cause* the variable y_r.

If there is at least one $a_{rs,i} \neq 0$ then y_s is said to *Granger-cause* y_r.

The presence of nonzero off-diagonal elements in the error covariance matrix Ω signals the presence of simultaneity.

The presence of $\sigma_{rs} \neq 0$ is sometimes referred to as instantaneous causality.

13.4 Determining The Lag Order of a Vector Autoregression

A traditional way of selecting the lag order is to use a sequential testing procedure.

Consider the VAR model with error covariance matrix $\mathbf{\Omega}_p = E\left(\mathbf{u}_t \mathbf{u}_t'\right)$, where a p subscript is included to emphasize that the matrix is related to a VAR(p). An estimate of this matrix is given by:

$$\hat{\mathbf{\Omega}}_p = (T-p)^{-1} \hat{\mathbf{U}}_p \hat{\mathbf{U}}_p'$$

where $\hat{\mathbf{U}}_p = \left(\hat{\mathbf{u}}_{p,1}', \ldots, \hat{\mathbf{u}}_{p,n}'\right)'$ is the matrix of residuals obtained by OLS estimation of the VAR(p), $\hat{\mathbf{u}}_{p,r} = \left(\hat{u}_{r,p+1}, \ldots, \hat{u}_{r,T}\right)'$ being the residual vector from the rth equation (noting that with a sample of size T, p observations will be lost through lagging).

A likelihood ratio (LR) statistic for testing the order p against the order m, m , p, is:

$$LR(p,m) = (T-np) \log \left(\frac{\left|\hat{\mathbf{\Omega}}_m\right|}{\left|\hat{\mathbf{\Omega}}_p\right|} \right) \sim \chi^2_{n^2(p-m)}$$

13.5 Variance Decompositions and Innovation Accounting

Vector moving average representation (VMA):

Suppose that the VAR is written in lag operator form as

$$\mathbf{A}(B)\mathbf{y}_t = \mathbf{u}_t$$

where, as in

$$\mathbf{A}(B) = \mathbf{I}_n - \mathbf{A}_1 B - \cdots - \mathbf{A}_p B^p$$

is a matrix polynomial in B. Analogous to the univariate case, the (infinite order) VMA representation is

$$\mathbf{y}_t = \mathbf{A}^{-1}(B)\mathbf{u}_t = \boldsymbol{\Psi}(B)\mathbf{u}_t = \mathbf{u}_t + \sum_{i=1}^{\infty} \boldsymbol{\Psi}_i \mathbf{u}_{t-i}$$

where

$$\boldsymbol{\Psi}_i = \sum_{i=1}^{i} \mathbf{A}_j \boldsymbol{\Psi}_{i-j} \quad \boldsymbol{\Psi}_0 = \mathbf{I}_n \quad \boldsymbol{\Psi}_i = \mathbf{0} \quad i < 0$$

this recursion being obtained by equating coefficients of B in $\boldsymbol{\Psi}(B)\mathbf{A}(B) = \mathbf{I}_n$

Impulse response function:

For sequence $\psi_{rs,1}, \psi_{rs,2}, \ldots$, where $\psi_{rs,i}$ is the rsth element of the matrix $\boldsymbol{\Psi}_i$.

Cholesky decomposition:

Define the lower triangular matrix $\mathbf{S}$ such that $\mathbf{S}\mathbf{S}' = \boldsymbol{\Omega}_p$ and define $\mathbf{v}_t = \mathbf{S}^{-1}\mathbf{u}_t$, then $E\left(\mathbf{v}_t \mathbf{v}_t'\right) = \mathbf{I}_n$

and the transformed errors v_t are orthogonal to each other.

In this case, the VMA representation can then be renormalized into the recursive form:

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \left(\boldsymbol{\Psi}_i \mathbf{S}\right)\left(\mathbf{S}^{-1}\mathbf{u}_{t-i}\right) = \sum_{i=0}^{\infty} \boldsymbol{\Psi}_i^O \mathbf{v}_{t-i}$$

where $\boldsymbol{\Psi}_i^O = \boldsymbol{\Psi}_i \mathbf{S}$ (so that $\boldsymbol{\Psi}_0^O = \boldsymbol{\Psi}_0 \mathbf{S}$ is lower triangular).

13.6 Structural Vector Autoregressions

The Cholesky decomposition can be written as $\mathbf{u}_t = \mathbf{S}\mathbf{v}_t$ with $\mathbf{S}\mathbf{S}' = \boldsymbol{\Omega}_p$ and $E\left(\mathbf{v}_t \mathbf{v}_t'\right) = \mathbf{I}_n$. A more general formulation is:

$$\mathbf{A}\mathbf{u}_t = \mathbf{B}\mathbf{v}_t$$

so that:

$$\mathbf{B}\mathbf{B}' = \mathbf{A}\boldsymbol{\Omega}_p \mathbf{A}'$$

Since both $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ matrices, they contain $2n^2$ elements, but the symmetry of the matrices on either side of (13.8) imposes $n(n+1)/2$ restrictions. A further $2n^2 - n(n+1)/2 = n(3n-1)/2$ restrictions, at least, must then be imposed to complete the identification of $\mathbf{A}$ and $\mathbf{B}$.

Chapter 14: Error Correction, Spurious Regressions, and Cointegration

14.1 The Error Correction Form of an Autoregressive Distributed Lag Model

Error-correction model (ECM):

The simplest case of the ARDL (autoregressive distributed lag) model is the ARDL(1, 1):

$$y_t = \beta_0 + \phi y_{t-1} + \beta_{1,0} x_t + \beta_{1,1} x_{t-1} + a_t$$

Suppose now that we recast this ARDL by writing it as:

$$\nabla y_t = \beta_0 - (1-\phi) y_{t-1} + (\beta_{1,0} + \beta_{1,1}) x_{t-1} + \beta_{1,0} \nabla x_t + a_t$$

or

$$\nabla y_t = \beta_{1,0} \nabla x_t - (1-\phi)\left(y_{t-1} - \frac{\beta_0}{1-\phi} - \frac{\beta_{1,0}+\beta_{1,1}}{1-\phi} x_{t-1} \right) + a_t$$

i.e., as

$$\nabla y_t = \beta_{1,0} \nabla x_t - (1-\phi)\left(y_{t-1} - \theta_0 - \theta_1 x_{t-1} \right) + a_t$$

is known as ECM.

If the parameters of the equilibrium relationship are unknown, then they may be estimated either by using nonlinear least squares on the model or by expressing the ECM as:

$$\nabla y_t = \beta_0 + \beta_{1,0} \nabla x_t + \gamma\left(y_{t-1} - x_{t-1} \right) + \delta x_{t-1} + a_t$$

It may readily be extended to the general $\mathrm{ARDL}\left(p, s_1, \ldots, s_M\right)$ model. Denoting the error correction as

$$ec_t = y_t - \theta_0 - \sum_{j=1}^{M} \theta_j x_{j,t}$$

and be generalized to

$$\begin{aligned} \nabla y_t = & \beta_0 - \phi(1) ec_{t-1} + \phi^*(B) \nabla y_{t-1} + \sum_{j=1}^{M} \tilde{\beta}_j(B) \nabla x_{j,t-1} \\ & + \sum_{i=1}^{M} \beta_j(B) \nabla x_{j,t} + a_t \end{aligned}$$

where

$$\phi^*(B) = \sum_{i=1}^{p} \phi_i B^i = \phi(B) - 1$$

14.2 Spurious Regressions

Data generation process (DGP):

$$\begin{aligned} y_t &= \phi y_{t-1} + u_t \quad u_t \sim \text{ i.i.d. } \left(0, \sigma_u^2\right) \\ x_t &= \phi^* x_{t-1} + v_t \quad v_t \sim \text{ i.i.d. } \left(0, \sigma_v^2\right) \end{aligned}$$

$$E\left(u_t v_s\right) = 0 \text{ for all } t, s \quad E\left(u_t u_{t-k}\right) = E\left(v_t v_{t-k}\right) = 0 \text{ for all } k \neq 0$$

i.e., that y_t and x_t are uncorrelated first-order autoregressive processes. Since x_t neither affects or is affected by y_t, it should be hoped that the coefficient β_1 in the regression model

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

would converge in probability to zero, reflecting the lack of any relationship between the two series, as would the R^2 statistic from this regression

14.3 Error Correction and Cointegration

Cointegration:

> 如果所考虑的时间序列具有相同的单整阶数，且某种线性组合协整向量）使得组合时间序列的单整阶数降低，则称这些时间序列之间存在显著的协整关系。也就是说，k 维向量 $Y_t = (y_{1t,}y_{2t}, \ldots, y_{kt})$ 的分量间被称为d, b阶协整，记为$Y_t \sim CI(d, b)$，如果满足：
>
> - y_{1t}，y_{2t}，…，y_{kt}都是 d 阶单整的，即$Y_t \sim I(d)$，要求 Y_t 的每个分量 $Y_{it} \sim I(d)$；
> - 存在非零向量 $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$，使得$\beta' Y_t \sim I(d, b)$，$0 < b \le d$，简称 Y_t 是协整的，向量β又称为协整向量。
>
> **协整关系存在的条件是**：只有当两个变量的时间序列{x}和{y}是同阶单整序列即I(d)时，才可能存在协整关系(这一点对多变量协整并不适用)。因此在进行y和x两个变量协整关系检验之前，先用ADF单位根检验对两时间序列{x}和{y}进行平稳性检验。平稳性的常用检验方法是图示法与单位根检验法。

14.4 Testing for Cointegration

Conditional error correction (CEC) form of an ARDL model:

$$\begin{aligned}\nabla y_t =& \alpha_0 + \alpha_1 t - \phi(1) y_{t-1} + \sum_{j=1}^{M} \beta_j(1) x_{j,t-1} \\ &+ \phi^*(B) \nabla y_{t-1} + \sum_{j=0}^{M} \gamma_j(B) \nabla x_{j,t} + a_t\end{aligned}$$

where

$$\gamma_j(B) = \beta_j(1) + \tilde{\beta}_j(B)$$

14.5 Estimating Cointegrating Regressions

Dynamic OLS (DOLS): deals with these problems by including leads and lags of $\nabla x_{j,t}$, and possibly lags of ∇y_t, as additional regressors in CEC so that standard OLS may continue to be used, i.e.,

$$y_t = \beta_0 + \sum_{j=1}^{M} \beta_{j,t} x_{j,t} + \sum_{i=1}^{p} \gamma_i \nabla y_{t-i} + \sum_{j=1}^{M} \sum_{i=-p_1}^{p_2} \delta_{j,i} \nabla x_{j,t-i} + e_t$$

An estimate of the cointegrating relationship is also provided by the error correction term in the appropriate form of the CEC. If there is no intercept or trend in the CEC then

$$ec_t = y_t - \sum_{j=1}^{M} \frac{\beta_j(1)}{\phi(1)} x_{j,t}$$

will provide estimates of the cointegrating parameters.

Chapter 15: Vector Autoregressions With Integrated Variables, Vector Error Correction Models, and Common Trends

15.1 Vector Autoregressions With Integrated Variables

There is a direct link between the coefficient matrices of

$$\boldsymbol{A}(B)\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{u}_t$$
$$\nabla \boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\Phi}(\mathbf{B})\nabla \boldsymbol{y}_{t-1} + \Pi \boldsymbol{y}_{t-1} + \boldsymbol{u}_t$$

that

$$\boldsymbol{A}_p = -\boldsymbol{\Phi}_{p-1}$$
$$\boldsymbol{A}_i = \boldsymbol{\Phi}_i - \boldsymbol{\Phi}_{i-1}, \quad i = 2, \ldots, p-1$$
$$\boldsymbol{A}_1 = \boldsymbol{\Pi} + \boldsymbol{I_n} + \boldsymbol{\Phi}_1$$

15.2 Vector Autoregressions With Cointegrated Variables

Vector error correction model (VECM):

$$\nabla \boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\Phi}(B)\nabla \boldsymbol{y}_{t-1} + \boldsymbol{\beta}\boldsymbol{e}_{t-1} + \boldsymbol{u}_t$$

where $\boldsymbol{e}_t = \boldsymbol{\alpha}'\boldsymbol{y_t}$ contains the r stationary **error corrections**. This is known as **Granger’s Representation Theorem** and is clearly the multivariate extension and generalization of generalized ARDL.

15.3 Estimation of Vector Error Correction Models and Tests of Cointegrating Rank

Reduced rank regression:

> 降秩回归是一种用于降维的多变量线性回归，适用于存在多个相关性较强的连续型因变量的情况。其在传统多变量线性回归的假设基础上，对回归系数矩阵或预测值矩阵的秩加以限制。假设自变量个数为p（即自变量组：$x_1, x_2, \ldots, x_p$），因变量个数为q（即因变量组：$y_1, y_2, \ldots, y_q$），样本量为n，自变量矩阵$X(n \times p)$与因变量矩阵 $Y(n \times p)$均满秩，且$n > p \geq q$。传统的普通最小二乘估计（ordinary least squares，OLS）的目标是使残差平方和最小，即求解回归系数矩阵$B(p \times q)$，使得$L = ||Y - XB||^2$最小。其中预测值矩阵用$\hat{Y}$表示，即$\hat{Y} = X\hat{B}$。
>
> 降秩回归方法对回归系数矩阵$B(p \times q)$的秩进行限制，要求在B的秩m≤q的前提下，使残差平方和最小。这等同于求解m个不存在共线性的自变量线性组合（即m个降秩回归因子），来解释因变量组的变异。该求解过程可以通过对普通最小二乘法得到的预测值矩阵$\hat{Y}$做奇异值分解（或基于其协方差矩阵做主成分分析）来实现，即$\hat{Y} = U\sum V^T$。m个最优线性组合的估计值由提取出的前m个特征向量（V_m）获得，即$\hat{Y}V_m$，而降秩回归的预测值矩阵$\hat{Y}_r = \hat{Y}V_mV_m^T$。同样的，降秩回归的回归系数矩阵的估计为$\hat{B}_r = \hat{B}V_mV_m^T$。从主成分分析的角度来看，降秩回归所得的m个自变量线性组合就是OLS估计出的预测值组的前m个主成分。

> 特别的，在规定m=q的情况下，降秩回归能提取出q个自变量线性组合但不能起到降维的作用，此时其优点是这q个变量相互线性独立。而在规定m < q的情况下，对秩次的这一限制则能体现降秩回归的降维作用。在现实情况下，使因变量组的预测残差平方和最小的秩次数为q，因为每减少一个秩次，预测变量的数量（即可以使用的预测信息）就会减少，对因变量组的拟合效果就会变差。因此降秩回归中对秩次m的选择是在降维和拟合度之间的取舍与平衡。此外，降秩回归是多变量统计方法，如果只有一个因变量（q=1），降秩回归的结果就等同于一般线性回归的结果。

15.4 Structural Vector Error Correction Models

Structural VECM:

$$\boldsymbol{\Gamma}_0 \nabla \boldsymbol{y}_t=\sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \nabla \boldsymbol{y}_{t-i}+\boldsymbol{\Theta} \boldsymbol{\alpha}^{\prime} \boldsymbol{y}_{t-1}+\boldsymbol{v}_t$$

which is related to the “reduced form” VECM

$$\nabla \boldsymbol{y}_t=\sum_{i=1}^{\boldsymbol{p}-1} \boldsymbol{\Phi}_i \nabla \boldsymbol{y}_{t-i}+\boldsymbol{\beta} \boldsymbol{\alpha}^{\prime} \boldsymbol{y}_{t-1}+\boldsymbol{u}_t$$

through

$$\begin{gathered}
\boldsymbol{\Gamma}_i=\boldsymbol{\Gamma}_0 \boldsymbol{\Phi}_i \quad i=1, \ldots, p-1 \\
\boldsymbol{\Gamma}_0 \boldsymbol{\beta}=\boldsymbol{\Theta} \quad \boldsymbol{\nu}_t=\boldsymbol{\Gamma}_0 \boldsymbol{u}_t
\end{gathered}$$

so that

$$E\left(\boldsymbol{\nu}_t \boldsymbol{\nu}_t^{\prime}\right)=\boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_p \boldsymbol{\Gamma}_0^{\prime}$$

15.5 Causal Testing in Vector Error Correction Models

Consider a “fully partitioned” form of the marginal VECM

$$\begin{aligned}
\nabla \boldsymbol{x}_t &=\boldsymbol{c}_1+\sum_{i=1}^{p-1} \boldsymbol{\Phi}_{11, i} \nabla \boldsymbol{x}_{t-i}+\sum_{i=1}^{p-1} \boldsymbol{\Phi}_{12, i} \nabla z_{t-i}+\boldsymbol{\beta}_1 \boldsymbol{\alpha}_1^{\prime} \boldsymbol{x}_{t-1}+\boldsymbol{\beta}_1 \boldsymbol{\alpha}_2^{\prime} z_{t-1}+\boldsymbol{u}_{1, t} \\
\nabla \boldsymbol{z}_t &=\boldsymbol{c}_2+\sum_{i=1}^{p-1} \boldsymbol{\Phi}_{21, i} \nabla \boldsymbol{x}_{t-i}+\sum_{i=1}^{p-1} \boldsymbol{\Phi}_{22, i} \nabla z_{t-i}+\boldsymbol{\beta}_2 \boldsymbol{\alpha}_1^{\prime} \boldsymbol{x}_{t-1}+\boldsymbol{\beta}_2 \boldsymbol{\alpha}_2^{\prime} z_{t-1}+\boldsymbol{u}_{1, t}
\end{aligned}$$

where now

$$\boldsymbol{\Phi}_i=\left[\begin{array}{ll}
\boldsymbol{\Phi}_{11, i} & \boldsymbol{\Phi}_{12, i} \\
\boldsymbol{\Phi}_{21, i} & \boldsymbol{\Phi}_{22, i}
\end{array}\right] \quad \boldsymbol{\alpha}^{\prime}=\left[\begin{array}{ll}
\boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2
\end{array}\right]^{\prime}$$

The hypothesis that z does not Granger-cause x may then be formalized as

$$\mathcal{H}_0: \boldsymbol{\Phi}_{12,1}=\cdots=\boldsymbol{\Phi}_{12, p-1}=\mathbf{0}, \quad \boldsymbol{\beta}_1 \boldsymbol{\alpha}_2^{\prime}=\mathbf{0}$$

The second part of $\mathcal{H}_0$, which is often referred to as “long-run noncausality,” involves a nonlinear function of the α and β coefficients and this complicates testing considerably.

15.6 Impulse Response Asymptotics in Nonstationary VARs

Impulse responses for nonstationary VARs should, therefore, not be computed from an unrestricted levels VAR. Since knowing the number of unit roots in the system is necessary for obtaining accurate estimates, it is important that the cointegrating rank is selected by a consistent method that works well in practice.

15.7 Vector Error Correction Model-X Models

A straightforward extension of the CVAR/VECM model is to include a vector of Ið Þ 0 exogenous variables, w_t say, which may enter each equation:

$$\nabla \boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{d}t + \sum_{i=1}^{p-1} \boldsymbol{\Phi}_i \nabla \boldsymbol{y}_{t-i} + \boldsymbol{\beta}\boldsymbol{\alpha}' \boldsymbol{y}_{t-1} + \boldsymbol{\Lambda} \boldsymbol{w}_t + \boldsymbol{u}_t$$

Estimation and testing for cointegrating rank remain exactly as before, although critical values of tests may be affected.

15.8 Common Trends and Circles

Consider

$$\boldsymbol{y}_t = \boldsymbol{b}_0 + \boldsymbol{C}(1)\boldsymbol{s}_t + \boldsymbol{C}^*(B)\boldsymbol{u}_t = \boldsymbol{C}(1)\left(\boldsymbol{c} + \boldsymbol{s}_t\right) + \boldsymbol{C}^*(B)\boldsymbol{u}_t$$

If there is cointegration, then as we have seen, $C(1)$ is of reduced rank $h = n - r$ and can be written as the product $\rho\delta'$, where both matrices are of rank h. On defining

$$\boldsymbol{\tau}_t = \boldsymbol{\delta}'\left(\boldsymbol{c} + \boldsymbol{s}_t\right) \quad \boldsymbol{c}_t = \boldsymbol{C}^*(B)\boldsymbol{u}_t$$

y_t can then be expressed in the “common trends” representation of Stock and Watson (1988):

$$\begin{gathered}\boldsymbol{y}_t = \boldsymbol{\rho}\boldsymbol{\tau}_t + \boldsymbol{c}_t \\ \boldsymbol{\tau}_t = \boldsymbol{\tau}_{t-1} + \boldsymbol{\delta}'\boldsymbol{u}_t\end{gathered}$$

This representation expresses $\boldsymbol{y}_t$ as a linear combination of $h = n - r$ random walks, these being the common trends τ_t, plus some stationary “transitory” components c_t. In fact, the transformed y_t may be regarded as a multivariate extension of the Beveridge-Nelson decomposition.

In the same way that common trends appear in y_t when $C(1)$ is of reduced rank, common cycles appear if $C^*(B)$ is of reduced rank.

Chapter 16: Compositional and Count Time Series

16.1 Constrained Time Series

Two examples of these types of series are considered in this chapter.

- compositional time series in which a group of series are defined as shares of a whole, so that they must be positive fractions that sum to unity
- “count” time series that can only take on positive, and typically low, integer values.

16.2 Modeling Compositional Data

A compositional data set is one in which the T observations on $D = d+1$ variables, written in matrix form as

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D} \\ \vdots & \vdots & & \vdots \\ x_{T,1} & x_{T,2} & \cdots & x_{T,D} \end{bmatrix} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_D]$$

where $\boldsymbol{x}_i = (x_{1,i}, x_{2,i}, \ldots, x_{T,i})', \quad i = 1, 2, \ldots, D$ are such that $x_{t,i} > 0$ and $x_{t,1} + x_{t,2} + \cdots + x_{t,D} = 1, t = 1, 2, \ldots, T$, that is, $\boldsymbol{x}_i > \boldsymbol{0}$ and $x_1 + x_2 + \cdots + x_D = \iota$, where $\iota = [1 \quad 1 \quad \cdots \quad 1]'$ is a $T \times 1$ unit vector. The sub-matrix $\boldsymbol{X}^{(d)} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_d]$ then lies in the **d-dimensional simplex $\mathcal{S}^d$ embedded in D-dimensional real space** with

$$\boldsymbol{x}_D = \boldsymbol{\iota} - \sum_{i=1}^{d} \boldsymbol{x}_i$$

being the vector of 'fill-up' values and $\boldsymbol{X} = \left[\boldsymbol{X}^{(d)} \quad \boldsymbol{x}_D\right]$

Additive log-ratio transformation:

> $$\begin{aligned} \boldsymbol{Y} &= [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_d] = a_d\left(\boldsymbol{X}^{(d)}\right) \\ &= \left[\log\left(\frac{\boldsymbol{x}_1}{\boldsymbol{x}_D}\right) \quad \log\left(\frac{\boldsymbol{x}_2}{\boldsymbol{x}_D}\right) \quad \cdots \quad \log\left(\frac{\boldsymbol{x}_d}{\boldsymbol{x}_D}\right)\right] \end{aligned}$$
>
> The inverse transformation, known as the additive-logistic, is
>
> $$\begin{aligned} \boldsymbol{X}^{(d)} &= a_d^{-1}(\boldsymbol{Y}) = \left[\frac{\exp(\boldsymbol{y}_1)}{\boldsymbol{y}} \quad \frac{\exp(\boldsymbol{y}_2)}{\boldsymbol{y}} \quad \cdots \quad \frac{\exp(\boldsymbol{y}_d)}{\boldsymbol{y}}\right] \\ \boldsymbol{x}_D &= \frac{1}{\boldsymbol{y}} \end{aligned}$$
>
> where
>
> $$\boldsymbol{y} = 1 + \sum_{i=1}^{d} \exp\left(\boldsymbol{y}_i\right)$$

Centered log-ratio transformation:

> $$\boldsymbol{Z} = c_d\left(\boldsymbol{X}^{(d)}\right) = \left[\log\left(\frac{\boldsymbol{x}_1}{g(\boldsymbol{X})}\right) \quad \log\left(\frac{\boldsymbol{x}_2}{g(\boldsymbol{X})}\right) \quad \cdots \quad \log\left(\frac{\boldsymbol{x}_D}{g(\boldsymbol{X})}\right)\right]$$
>
> where
>
> $$g(\boldsymbol{X}) = \begin{bmatrix} \left(x_{1,1} \times x_{1,2} \times \cdots \times x_{1,D}\right)^{1/D} \\ \left(x_{2,1} \times x_{2,2} \times \cdots \times x_{2,D}\right)^{1/D} \\ \vdots \\ \left(x_{T,1} \times x_{T,2} \times \cdots \times x_{T,D}\right)^{1/D} \end{bmatrix}$$
>
> is the vector of geometric means. Unfortunately, this has the disadvantage of introducing a non-singularity since $Z_\iota = 0$

16.3 Forecasting Compositional Time Series

Let us now denote the tth rows of X and Y as X_t and Y_t; respectively, and let us assume that an h-step ahead forecast of Y_{t+h}, which may not yet be observed, is available. This may be denoted $Y_t(h)$ with covariance matrix $\sum_t(h)$. Since Y_t is multivariate normal, such forecasts may have been obtained from a wide variety of multivariate models.

16.4 IN-AR(1) BENCHMARK MODEL

Integer-valued ARMA (IN-ARMA) models:

> It provides an interesting class of discrete valued processes that are able to specify not only the dependence structure of the series of counts, but also enable a choice to be made between a wide class of (discrete) marginal distributions.

"Benchmark" IN-AR(1) process:

> $$x_t = a \circ x_{t-1} + w_t$$
>
> where the $x_t, \quad t = 1, 2, \ldots$, take on values in the set of nonnegative integers, $\mathcal{N} = \{0, 1, 2, \ldots\}$.
>
> It is assumed $0 \leq a < 1$ and that w_t is a sequence of i.i.d. discrete random variables with mean $\mu_w > 0$ and variance $\sigma_w^2 > 0$: w_t is assumed to be stochastically independent of x_{t-1} for all t.

Binomial thinning operation:

> The process is stationary and the discreteness of x_t is ensured by
>
> $$a \circ x_{t-1} = \sum_{i=1}^{x_{t-1}} y_{i,t-1}$$
>
> where the $y_{i,t-1}$ are assumed to be i.i.d. Bernoulli random variables with
>
> $$\begin{gathered} P\left(y_{i,t-1} = 1\right) = a \\ P\left(y_{i,t-1} = 0\right) = 1 - a \end{gathered}$$

The unconditional moments of x_t are

$$E\left(x_t\right) = \frac{\mu_w}{(1-a)} \quad V\left(x_t\right) = \frac{\left(a\mu_w + \sigma_w^2\right)}{\left(1-a^2\right)}$$

while the conditional moments of x_t are

$$E\left(x_t \mid x_{t-1}\right) = a x_{t-1} + \mu_w \quad V\left(x_t \mid x_{t-1}\right) = a(1-a) x_{t-1} + \sigma_w^2$$

so that both are linear in x_{t-1}.

16.5 Estimation of Integer-valued ARMA Models

The "bias-corrected Yule-Walker" estimate:

Using a "bias-corrected" first-order sample autocorrelation to estimate a:

$$\hat{a} = \frac{1}{T-3}(\operatorname{Tr}_1 + 1)$$

The estimate of λ is then based on the moment condition $E\left(x_t\right) = \lambda/(1-a)$:

$$\hat{\lambda} = (1-\hat{a})\bar{x}$$

16.6 Testing for Serial Dependence in Count Time Series

Before fitting a member of the IN-ARMA class of models it is important to establish the nature of the serial dependence, if any, in a time series of counts.

Three tests:

- $$S^* = \sqrt{T} r_1 \sim N(0,1)$$

  under the null hypothesis of i.i.d. Poisson random variables, with a one-sided test being used that rejects the null for large values of the statistic.

- $$Q_{\text{acf}}(1) = \frac{\hat{r}_2^2 \left(\sum_{t=1}^{T} \left(x_t - \bar{x}\right)^2 \right)^2}{\sum_{t=3}^{T} \left(x_t - \bar{x}\right)^2 \left(x_{t-2} - \bar{x}\right)^2}$$

- $$Q_{\text{pacf}}(1) = \frac{\hat{\phi}_2^2 \left(\sum_{t=1}^{T} \left(x_t - \bar{x}\right)^2 \right)^2}{\sum_{t=3}^{T} \left(x_t - \bar{x}\right)^2 \left(x_{t-2} - \bar{x}\right)^2}$$

  where $\hat{\phi}_2$ is the second-order sample partial autocorrelation.

Under the i.i.d. Poisson null hypothesis, these statistics are asymptotically distributed as $\chi^2(1)$.

16.7 Forecasting Counts

$$\begin{gathered} f_{T,1} = a x_T + \mu_w \\ f_{T,2} = a f_{T,1} + \mu_w = a^2 x_T + (1+a)\mu_w \\ f_{T,h} = a^h x_T + \left(1 + a + a^2 + \cdots + a^{h-1}\right)\mu_w \end{gathered}$$

Since $0 \leq a < 1$ the forecasts converge as $h \rightarrow \infty$ to the unconditional mean $E\left(x_t\right) = \mu_w / (1-a)$.

16.8 Intermittent and Nonnegative Time Series

Intermittent:

When a count series contains many zeros, it is sometimes referred to as being intermittent.

Chapter 17: State Space Models

17.1 Formulating State Space Models

State space form (SSF):

Many time series models can be cast in SSF, and this enables a unified framework of analysis to be presented within which, for example, the differences and similarities of the alternative models may be assessed.

The state space model for a univariate time series x_t consists of both a measurement equation (alternatively known as the signal or observation equation) and a transition equation (alternatively state equation.

A popular version has the measurement equation taking the form:

$$x_t = z_t' \boldsymbol{\alpha}_t + d_t + \varepsilon_t \quad t = 1, 2, \ldots, T$$

Here z_t is an $m \times 1$ vector, d_t is a scalar, and ε_t is a serially uncorrelated error with $E(\varepsilon_t) = 0$ and $V(\varepsilon_t) = h_t$, state vector α_t which is generated by

$$\boldsymbol{\alpha}_t = \boldsymbol{T}_t \boldsymbol{\alpha}_{t-1} + \boldsymbol{c}_t + \boldsymbol{R}_t \boldsymbol{\eta}_t$$

The specification of the state space system is completed by two further assumptions:

- The initial state α_0 has mean vector $E(\boldsymbol{\alpha}_0) = \boldsymbol{a}_0$ and covariance matrix $V(\boldsymbol{\alpha}_0) = \boldsymbol{P}_0$;
- The errors ε_t and η_t are uncorrelated with each other in all time periods and uncorrelated with the initial state, that is,

$$E\left(\varepsilon_t \boldsymbol{\eta}_s'\right) = \mathbf{0} \text{ for all } s, t = 1, \ldots, T$$

and

$$E\left(\varepsilon_t \boldsymbol{\alpha}_0'\right) = \mathbf{0} \quad E\left(\boldsymbol{\eta}_t \boldsymbol{\alpha}_0'\right) = \mathbf{0} \quad \text{ for all } t = 1, \ldots, T$$

The variables z_t, d_t, and h_t in the measurement equation and T_t, c_t, R_t, and Q_t in the transition equation are referred to generically as the **system matrices**.

17.2 The Kalman Filter

Consider the state space model of (17.1) and (17.2). Let $\boldsymbol{a}_{t-1}$ be the optimal estimator of $\boldsymbol{\alpha}_{t-1}$ based on observations up to and including x_{t-1}, that is, $\boldsymbol{a}_{t-1} = E_{t-1}(\boldsymbol{\alpha}_{t-1} \mid \boldsymbol{x}_{t-1})$, where $\boldsymbol{x}_{t-1} = \{x_{t-1}, x_{t-2}, \ldots, x_1\}$, and let

$$\boldsymbol{P}_{t-1} = E(\boldsymbol{\alpha}_{t-1} - \boldsymbol{a}_{t-1})(\boldsymbol{\alpha}_{t-1} - \boldsymbol{a}_{t-1})'$$

be the $m \times m$ covariance matrix of the estimation error. Given $\boldsymbol{a}_{t-1}$ and $\boldsymbol{P}_{t-1}$, the optimal estimators of $\boldsymbol{\alpha}_t$ and $\boldsymbol{P}_t$ are given by:

$$\boldsymbol{a}_{t|t-1} = \boldsymbol{T}_t \boldsymbol{a}_{t-1} + \boldsymbol{c}_t$$

and

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{T}_t \boldsymbol{P}_{t-1} \boldsymbol{T}_t' + \boldsymbol{R}_t \boldsymbol{Q}_t \boldsymbol{R}_t'$$

These two recursions are known as the prediction equations. Once the new observation x_t becomes available, the estimator of $\boldsymbol{\alpha}_t, \boldsymbol{a}_{t|t-1}$, can be updated. The updating equations are:

$$\boldsymbol{a}_t = \boldsymbol{a}_{t|t-1} + \boldsymbol{P}_{t|t-1} z_t f_t^{-1} \left(x_t - z_t' \boldsymbol{a}_{t|t-1} - d_t\right)$$

and

$$\boldsymbol{P}_t = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1} z_t f_t^{-1} z_t' \boldsymbol{P}_{t|t-1}$$

where

$$f_t = z_t' \boldsymbol{P}_{t|t-1} z_t + h_t$$

Taken together, Eqs. (17.4-17.8) make up the **Kalman filter**. These equations may also be written as

$$\boldsymbol{a}_{t+1|t} = \left(\boldsymbol{T}_{t+1} - \boldsymbol{K}_t z_t'\right) \boldsymbol{a}_{t|t-1} + \boldsymbol{K}_t x_t + \boldsymbol{c}_{t+1} - \boldsymbol{K}_t d_t$$

where the $m \times 1$ gain vector $\boldsymbol{K}_t$ is given by

$$\boldsymbol{K}_t = \boldsymbol{T}_{t+1}\boldsymbol{P}_{t|t-1} z_t f_t^{-1}$$

The recursion for the error covariance matrix, known as the **Riccati equation**, is

$$\boldsymbol{P}_{t+1|t} = \boldsymbol{T}_{t+1}\left(\boldsymbol{P}_{t|t-1} - f_t^{-1}\boldsymbol{P}_{t|t-1}\boldsymbol{z}_t z_t' \boldsymbol{P}_{t|t-1}\right)\boldsymbol{T}_{t+1}' + \boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}\boldsymbol{R}_{t+1}'$$

17.3 MLE and The Prediction Error Decomposition

Hyperparameters of the SSF:

Hyperparameters may be estimated by ML, the classical theory of which is based on the T observations $x_1, \ldots, x_T$ being i.i.d. This allows the joint density function of the observations to be written as:

$$\mathcal{L}(\boldsymbol{x} : \boldsymbol{\psi}) = \prod_{t=1}^{T} p\left(x_t\right)$$

where $\boldsymbol{x}' = \left(x_1, \ldots, x_T\right)$ and $p\left(x_t\right)$ is the probability density function of x_t. Once the observations have become available.

If the measurement equation is written as:

$$x_t = z_t' \boldsymbol{a}_{t|t-1} + z_t'\left(\boldsymbol{\alpha}_t - \boldsymbol{a}_{t|t-1}\right) + d_t + \varepsilon_t$$

then the conditional distribution of xt is normal with mean

$$E_{t-1}\left(x_t\right) = \hat{x}_{t|t-1} = z_t' \boldsymbol{a}_{t|t-1} + d_t$$

and variance f_t. The likelihood function can then be written as

$$\log \mathcal{L} = \ell = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T} f_t - \frac{1}{2}\sum_{t=1}^{T} \nu_t^2 / f_t$$

where $\nu_t = x_t - \hat{x}_{t|t-1}$ is the **prediction error**, so that $\log \mathcal{L}$ is also known as the **prediction error decomposition** form of the likelihood function.

17.4 Prediction and Smoothing

Smoothing:

> While the aim of filtering is to find the expected value of the state vector, $\boldsymbol{\alpha}_t$, conditional on the information available at time t, that is, $\boldsymbol{a}_{t|t} = E\left(\boldsymbol{\alpha}_t \mid \boldsymbol{x}_t\right)$, the aim of smoothing is to take account of the information available after time t. This will produce the smoothed estimator $\boldsymbol{a}_{t|T} = E\left(\boldsymbol{\alpha}_t \mid \boldsymbol{x}_T\right)$ and, since it is based on more information than the filtered estimator, it will have an MSE which cannot be greater than that of the filtered estimator.

Fixed-interval smoothing:

> This is an algorithm that consists of a set of recursions which start with the final quantities, a_T and P_T, given by the Kalman filter and work backward.These equations are:
>
> $$\boldsymbol{a}_{t|T} = \boldsymbol{a}_T + \boldsymbol{P}_t^*\left(\boldsymbol{a}_{t+1|T} - \boldsymbol{T}_{t+1}\boldsymbol{a}_t\right)$$
>
> and
>
> $$\boldsymbol{P}_{t|T} = \boldsymbol{P}_t + \boldsymbol{P}_t^*\left(\boldsymbol{P}_{t+1|T} - \boldsymbol{P}_{t+1|T}\right)\boldsymbol{P}_t^{*\prime}$$
>
> where

> $$\boldsymbol{P}_t^* = \boldsymbol{P}_t \boldsymbol{T}_{t+1}' \boldsymbol{P}_{t+1|t}^{-1} \quad t = T-1, \ldots, 1$$
>
> with $\boldsymbol{a}_{T|T} = \boldsymbol{a}_T$ and $\boldsymbol{P}_{T|T} = \boldsymbol{P}_T$.

17.5 Multivariate State Space Models

This development of state space models has been based on modeling a univariate time series x_t. The analysis may readily be extended to modeling the $N \times 1$ vector $\boldsymbol{X}_t$ of observed series by generalizing the measurement equation (17.1) to

$$\boldsymbol{X}_t = \boldsymbol{Z}_t \boldsymbol{\alpha}_t + \boldsymbol{d}_t + \varepsilon_t$$

where $\boldsymbol{Z}_t$ is an $N \times m$ matrix, $\boldsymbol{d}_t$ is an $N \times 1$ vector, and ε_t is an $N \times 1$ vector with $E(\varepsilon_t) = \mathbf{0}$ and $V(\varepsilon_t) = \boldsymbol{H}_t$, an $N \times N$ covariance matrix. The analysis then carries through with the necessary changes.