

Applied Time Series Analysis

Chapter 1 Time Series and Their Features

1.1 Time Series

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 $t = 1, 2, \dots, T$ will often be referred to as the **observation period**.

1.2 Feature 1: Autocorrelation

lag-k (sample) autocorrelation:

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

1.3 Feature 2: Seasonality

When a time series is observed at monthly or quarterly intervals an annual **seasonal pattern** is often an important feature.

1.4 Feature 3: Stationarity and Non-stationarity

A constant mean level is one, but not the only, condition for a series to be **stationarity**. If the mean level cannot be regarded as constant then a series is said to be **nonstationarity**.

1.5 Feature 4: Trends

Series with reasonably constant slopes appear to exhibit **linear trends**

1.6 Feature 5: Volatility

A second condition of **stationarity** is that of constant variance.

1.7 Feature 6: Common Features

Two or more time series may contain **common features**(like **co-integrate**).

1.8 Feature 7: Natural Constraints

Such as **compositional time series** and **count**

Chapter 2 Transforming Time Series

Prior to analyzing, statistically, an individual or a group of time series, it is often appropriate to transform the data, with an initial plot of the series often providing clues as to what transformation(s) to use. There are three general classes of transformations for time series—**distributional, stationarity inducing**, and **decompositional**—and these may often be combined to produce an appropriate variable to analyze

2.1 Distributional Transformations

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

$$f^{SP}(x_T, \lambda) = \begin{cases} (x_t^\lambda - 1)/\lambda & \text{if } \lambda = 0 \\ \log x_t & \text{if } \lambda \neq 0 \end{cases} \quad (2.1)$$

The restriction to positive values that is required by the Box-Cox transformation can be relaxed in several ways:

- the ***signed power*** transformation proposed by Bickel and Doksum (1981):

$$f^{GP}(x_T, \lambda) = (\operatorname{sgn}(x_t) |x_t^\lambda - 1|)/|\lambda| \quad \lambda > 0 \quad (2.2)$$

- the ***generalized power*** (GP) transformation suggested by Yeo and Johnson(2000):

$$f^{GP}(x_t, \lambda) = \begin{cases} ((x_t + 1)^\lambda - 1)/\lambda & x_t \geq 0, \lambda \neq 0 \\ \log(x_t + 1) & x_t \geq 0, \lambda = 0 \\ -((-x_t + 1)^{2-\lambda} - 1)/(2 - \lambda) & x_t < 0, \lambda \neq 2 \\ -\log(-x_t + 1) & x_t < 0, \lambda \neq 2 \end{cases} \quad (2.3)$$

- the ***inverse hyperbolic sine*** (IHS) transformation suggested by Burbidge et al. (1988) to deal with extreme values of either sign:

$$f^{IHS}(x_t, \lambda) = \frac{\sinh^{-1}(\lambda x_t)}{\lambda} = \log \frac{\lambda x_t + (\lambda^2 x_t^2 + 1)^{1/2}}{\lambda} \quad \lambda > 0 \quad (2.4)$$

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 $f(x_t, \lambda) = \mu_t + a_t$ 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 $\hat{\lambda}$ is then obtained by maximizing λ the concentrated log-likelihood function:
over where

$$\ell(\lambda) = C_f - \left(\frac{T}{2}\right) \sum_{t=1}^T \log \hat{a}_t^2 + D_f(x_t, \lambda)$$

$\hat{a}_t = f(x_t, \lambda) - \hat{\mu}_t$ 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:

$$\begin{aligned} D_f(x_t, \lambda) &= (\lambda - 1) \sum_{t=1}^T \log |x_t| && \text{for (2.1) and (2.2)} \\ &= (\lambda - 1) \sum_{t=1}^T \operatorname{sgn}(x_t) \log(|x_t| + 1) && \text{for (2.3)} \\ &= -\frac{1}{2} \sum_{t=1}^T \log(1 + \lambda^2 x_t^2) && \text{for (2.4)} \end{aligned}$$

2.2 Stationarity Inducing Transformations

lag operator $B : B^j x_t \equiv x_{t-j}$

first-difference : $\nabla x_t = x_t - x_{t-1} = x_t - Bx_t = (1 - B)x_t$

second-differences : $\nabla \nabla x_t = \nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2}$

two-period difference : $\nabla_2 x_t = 1 - B^2 x_t = x_t - x_{t-2}$

proportional or **percentage** changes: $\nabla x_t / x_{t-1}$ 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**.)

There is a useful relationship between the rate of change of a variable and its logarithm that is a more convenient transformation to work with and often worth bearing in mind, namely:

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

2.3 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 “permanent” movements from shorter-run, more “transitory,” fluctuations, that is, by separating the observations through a **decomposition**, generally of the form “data 5 fit 1 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 (1977). Tukey himself favored **running** or **moving medians** to do this, but **moving averages** (MAs) have become by far the most popular approach to smoothing a time series.

the $(2n + 1)$ -term **weighted** and **centered** MA [WMA($2n + 1$)] replaces x_t with

$$WMA_t(2n + 1) = \sum_{i=-n}^n \omega_i x_{t-i}$$

where the **weights** ω_i are restricted to sum to unity: $\sum_{i=-n}^n \omega_i = 1$ The weights are often symmetrically valued about the central weight ω_0 and because there is an odd number of weights in (2.9), $WMA_t(2n + 1)$ “matches up” with x_t , hence the use of the term “centered”.

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 X_t , being decomposed into trend, T_t , seasonal, S_t , and irregular, I_t , components. The decomposition can either be ***additive***:

$$X_t = T_t + S_t + I_t$$

or **multiplicative**

$$X_t = T_t \times S_t \times I_t$$

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 thus be random and hence unpredictable.

Chapter 3 ARMA Models for Stationary Time Series

3.1 Stochastic Processes and Stationarity

It is often useful to regard the observations x_1, x_2, \dots, x_T on the series x_t as a **realization** of a stochastic process.

Stationarity requires the process to be in a state of "statistical equilibrium."

A stochastic process is said to be **strictly stationary** if its properties are unaffected by a change of time origin, that is, the joint probability distribution at any set of times t_1, t_2, \dots, t_m must be the same as the joint probability distribution at $t_{1+k}, t_{2+k}, \dots, t_{m+k}$, where k is an arbitrary shift in time.

lag- k **autocovariance**:

$$\gamma_k = \text{Cov}(x_t, x_{t-k})$$

(**population**) **autocorrelation function (ACF)**:

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

ACF 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 and Autocorrelation

A fundamental theorem in time series analysis, known as **Wold's decomposition**, states that every weakly stationary, purely nondeterministic, stochastic process $x_t - \mu$ can be written as a linear combination (or linear **filter**) of a sequence of uncorrelated random variables. "Purely nondeterministic" means that any deterministic components have been subtracted from $x_t - \mu$. Such components are those that can be perfectly predicted from past values of themselves and examples commonly found are a (constant) mean, as is implied by writing the process as $x_t - \mu$, periodic sequences (e.g., sine and cosine functions), and polynomial or exponential sequences in t .

This linear filter representation is given by:

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

3.3 First-Order Autoregressive Processes

first-order autoregressive (AR(1)) process :

taking $\mu = 0$ (without loss of generality), choosing $\psi_j = \phi^j$:

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

or

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

or

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

3.4 First-Order Moving Average Processes

first-order moving average (MA(1)) process :

consider $\phi_1 = -\theta$ and $\phi_j = 0 \quad j \geq 2$:

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

or

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

3.5 General AR and MA Processes

1. general autoregressive model of order p (AR(p)):

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

or

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

kth partial autocorrelation is the coefficient ϕ_{kk} in the AR(k) process:

$$x_t = \phi_{k1} x_{t-1} + \phi_{k2} x_{t-2} + \cdots + \phi_{kk} x_{t-k} + a_t$$

2. general MA of order q (MA(q)) :

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

or

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

ACF:

$$\rho_k = \frac{-\theta_k + \theta_1 \theta_{k+1} + \cdots + \theta_{q-k} \theta_q}{1 + \theta_1^2 + \cdots + \theta_q^2} \quad k = 1, 2, \dots, q$$

$$\rho_k = 0 \quad k > q$$

3.6 Autoregressive-Moving Average Models

1. **first-order autoregressive-moving average** (ARMA(1,1)) process:

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

or

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

2. **general ARMA** models obtained by combining AR(p) and MA(q) processes:

$$x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

or

$$(1 - \phi_1 B - \cdots - \phi_p B^p)x_t = (1 - \theta_1 - \cdots - \theta_q B^q)a_t$$

which may be written more concisely as:

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

3.7 ARMA Model Building and Estimation

With the stationarity and (implicit) ergodicity assumptions, μ and σ_x^2 can be estimated by the sample mean and sample variance.

An estimate of ρ_k is then provided by the ***lag k sample autocorrelation***:

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

On the hypothesis that $x_t \sim WN(\mu, \sigma^2)$, then Ljung and Box (1978) show that:

$$Q(k) = T(T+2) \sum_{i=1}^k (T-i)^{-1} r_i^2 \stackrel{a}{\sim} \chi^2(k)$$

and this statistic may be used to assess whether an observed series departs significantly from white noise.

AIC and **BIC** to choose an appropriate model:

$$\begin{aligned} AIC(p, q) &= \log \hat{\sigma}^2 + 2(p+q)T^{-1} \\ BIC(p, q) &= \log \hat{\sigma}^2 + (p+q)T^{-1} \log T \end{aligned}$$

Chapter 4 ARIMA Models for Nonstationary Time Series

4.1 Nonstationarity

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

$$x_t = \mu_t + \epsilon_t$$

An alternative way of generating a nonstationary mean level is to employ ARMA models whose autoregressive parameters do not satisfy stationarity conditions.

4.2 ARIMA Processes

If the 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 $I(d)$.

4.3 ARIMA Modeling

Once the order of differencing d has been established, then since $w_t = \nabla^d x_t$ is, by definition, stationary, the ARMA model building techniques may be applied to the suitably differenced series.

Chapter 5 Unit Roots, Difference and Trend Stationarity, and Fractional Differencing

5.1 Determining the Order of Integration of a Time Series

- 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;
 1. the variance of x_t is finite and does not depend on t ;
 2. the innovation a_t has only a temporary effect on the value of x_t ;
 3. the expected length of time between crossings of $x = 0$ is finite, so that x_t fluctuates around its mean of zero;
 4. 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:
 1. the variance of x_t goes to infinity as t goes to infinity;
 2. an innovation a_t has a permanent effect on the value of x_t because x_t is the sum of all previous innovations: $x_t = \nabla^{-1} a_t = S a_t = \sum_{i=0}^{t-1} a_{t-i}$
 3. the expected time between crossings of $x = 0$ is infinite;
 4. the autocorrelations ρ_k for all k as t goes to infinity.

Sole reliance on the SACF can sometimes lead to problems of **overdifferencing**.

5.2 Testing 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 .

1. zero mean AR(1) process:

The OLS estimator of ϕ is given by:

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

2. AR(p)and ARMA(p,q) process:

a test of $\phi = 1$ can be constructed as

$$\tau_\mu = \frac{\hat{\phi}_T - 1}{\text{se}(\hat{\phi}_T)}$$

5.3 Trend Versus Difference Stationarity

difference stationary (DS):

$$\nabla x_t = \varepsilon_t$$

where $\varepsilon_t = \theta(B)a_t$, while the alternative is that x_t is stationary in levels

trend stationarity (TS):

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

5.4 Testing for More Than One Unit Root

If the null hypothesis of a unit root is not rejected, then it may be necessary to test whether the series contains a second unit root—in other words whether x_t is $I(2)$ and, thus, needs differencing twice to induce stationarity.

5.5 Other Approaches to Testing for a Unit Root

- Under a specific set of conditions placed upon a_t , known as ***weak dependency***, which are described in detail by Phillips (1987), the $\$$ statistic is modified to:

$$Z(\tau_\mu) = \tau_\mu (\hat{\sigma}_0 / \hat{\sigma}_\ell) - \frac{1}{2} (\hat{\sigma}_\ell^2 - \hat{\sigma}_0^2) / \Sigma_\ell$$

in which

$$\begin{aligned}\hat{\sigma}_0^2 &= T^{-1} \sum_{t=1}^T \hat{a}_t^2 \\ \hat{\sigma}_\ell^2 &= \hat{\sigma}_0^2 + 2T^{-1} \sum_{j=1}^{\ell} w_j(\ell) \left(\sum_{t=j+1}^T \hat{a}_t \hat{a}_{t-j} \right) \\ \Sigma_\ell^2 &= T^{-2} \hat{\sigma}_\ell^2 \sum_{t=2}^T (x_{t-1} - \bar{x}_{-1})^2 \quad \bar{x}_{-1} = (T-1)^{-1} \sum_{t=1}^{T-1} x_t\end{aligned}$$

- The DF-GLS and point optimal unit root tests of Elliott, Rothenberg, and Stock (ERS, 1996) were the initial pair of tests based on this approach, both employing the ***quasi-differences***:

$$d(x_t | \alpha) = \begin{cases} x_t & \text{if } t = 1 \\ x_t - \alpha x_{t-1} & \text{if } t > 1 \end{cases}$$

The ***point optimal*** test is the most powerful test of a unit root against a simple point alternative. If we define the residual sum of squares from the first-stage regression as $S(\bar{\alpha})$ then the point optimal test of the null $\phi = 1$ against the alternative $\phi = \bar{\phi} < 1$ is then defined as

$$P_\tau = \frac{S(\bar{\phi}) - \bar{\phi} S(1)}{\hat{\sigma}_\ell^2}$$

- Ng and Perron (2001) construct four further tests that are based on the GLS-detrended data x_t^d . It is useful to define the term

$$\kappa = T^{-2} \sum_{t=1}^{T-1} (x_t^d)^2$$

whereupon the test statistics are defined as

$$\begin{aligned}\text{MZ}^d(\phi) &= \frac{T^{-1} (x_T^d) - \hat{\sigma}_\ell^2}{2\kappa} \\ \text{MSB}^d &= \left(\frac{\kappa}{\hat{\sigma}_\ell^2} \right)^{1/2} \\ \text{MZ}_t^d &= \text{MZ}^d(\phi) \times \text{MSB}^d \\ \text{MP}_T^d(k) &= \frac{\bar{c}_k^2 \kappa - (\bar{c}_k - k) T^{-1} (x_T^d)^2}{\hat{\sigma}_\ell^2}\end{aligned}$$

4. The statistic proposed to test this null is the KPSS test (after Kwiatkowski et al., 1992), which is defined as

$$\eta_\tau = T^{-2} \sum_{t=1}^T \hat{S}_t^2 / \hat{\sigma}_{el}^2$$

Here

$$\hat{S}_t = \sum_{i=1}^t e_i \quad e_t = x_t - \hat{\beta}_0 - \hat{\beta}_1 t$$

5.6 Estimating Trends Robustly

This circularity of reasoning has prompted the development of trend function testing procedures that are robust, in the sense that, at least asymptotically, inference on the trend function is unaffected as to whether ε_t is $I(0)$ or $I(1)$.

What if it is not known whether ε_t is $I(0)$ or $I(1)$? Harvey, Leybourne, and Taylor (HLT, 2007) show that a weighted average of z_0 and z_1 , say

$$z_\lambda = (1 - \lambda(U, S))z_0 + \lambda(U, S)z_1$$

where U is a standard unit root test statistic, S is a standard trend-stationarity test statistic and

$$\lambda = \exp\left(-\kappa\left(\frac{U}{S}\right)^2\right)$$

5.7 Fractional Differencing and Long Memory

Many empirically observed time series, however, although appearing to satisfy the assumption of stationarity (perhaps after differencing), nevertheless seem to exhibit some dependence between distant observations that, although small, is by no means negligible. This may be termed **long range persistence** or **dependence**, although the term **long memory** is now popular.

The class of ARIMA processes may be extended to model this type of long-range persistence by relaxing the restriction to just integer values of d, so allowing **fractional differencing** within the class of AR-fractionally integrated-MA (ARFIMA) processes.

$$\begin{aligned} \nabla^d &= (1 - B)^d = \sum_{k=0}^{\infty} \frac{d!}{(d-k)!k!} (-B)^k \\ &= 1 - dB + \frac{d(d-1)}{2!} B^2 - \frac{d(d-1)(d-2)}{3!} B^3 + \dots \\ &\quad \text{for any real } d > -1 \end{aligned}$$

The long range persistence exhibited by fractionally differenced processes is often referred to as **long memory**.

5.8 Testing for Fractional Differencing

The “classic” approach to detecting the presence of long memory in a

time series is to use the **range over standard deviation** or **rescaled range** (R/S) statistic:

$$R_0 = \hat{\sigma}_0^{-1} \left[\max_{1 \leq i \leq T} \sum_{t=1}^i (x_t - \bar{x}) - \min_{1 \leq i \leq T} \sum_{t=1}^i (x_t - \bar{x}) \right] \quad \hat{\sigma}_0^2 = T^{-1} \sum_{t=1}^T (x_t - \bar{x})^2$$

modified by Lo (1991):

$$R_q = \hat{\sigma}_q^{-1} \left[\max_{1 \leq i \leq T} \sum_{t=1}^i (x_t - \bar{x}) - \min_{1 \leq i \leq T} \sum_{t=1}^i (x_t - \bar{x}) \right]$$

where

$$\hat{\sigma}_q^2 = \hat{\sigma}_0^2 \left(1 + \frac{2}{T} \sum_{j=1}^q w_{qj} r_j \right) w_{qj} = 1 - \frac{j}{q+1}, \quad q < T$$

An obvious approach to testing for fractional differencing is to construct tests against the null of either $d = 1$ or $d = 0$.

- Breitung and Hassler (2002), building upon the Lagrange Multiplier (LM) approach of Agiakloglou and Newbold (1994), show that a simple test of this null is the t-statistic testing $\phi = 0$ from the regression:

$$\nabla^d x_t = \phi x_{t-1}^* + a_t \text{ where}$$

where

$$x_{t-1}^* = \sum_{j=1}^{t-1} j^{-1} \nabla^d x_{t-j}$$

- Dolado et al. (2002) propose a similar DF regressin for testing the null that $x_t \sim I(d_0)$ against the alternative that $x_t \sim I(d_1)$, where d_0 and d_1 are real numbers. Their "FD-F" regression is

$$\nabla^{d_0} x_t = \phi \nabla^{d_1} x_{t-1} + a_t$$

in which $\nabla^{d_0} x_t$ and $\nabla^{d_1} x_{t-1}$ have been differenced according to their order of integration under the null and alternative hypothesis, respectively.

5.9 Estimating the Fractional Differencing Parameter

A popular and relatively simple estimator is the log-periodogram regression proposed by Geweke and Porter-Hudak (1983, GPH).

$$y_t = \nabla \hat{d} x_t = \sum_{k=0}^{t-1} \frac{\hat{d}!}{(\hat{d}-k)!k!} (-1)^k x_{t-k}$$

Chapter 6 Breaking and Nonlinear Trends

6.1 Breaking Trend Models

- ***level shift:***

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

$$\nabla x_t = \beta_0 + \mu \nabla \text{DU}_t^c + \varepsilon_t^* = \beta_0 + \mu \text{D}(\text{TB}^c)_t + \varepsilon_t^*$$

- ***segmented trend:***

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

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

- **combined:**

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

6.2 Breaking Trends and Unit Root Tests

Perron (1989) consequently extended the Dickey Fuller unit root testing strategy to ensure consistency against shifting trend functions by developing two asymptotically equivalent procedures:

- modified ADF regression

$$\tilde{x}_t^i = \tilde{\phi}^i \tilde{x}_{t-1}^i + \sum_{j=0}^k \gamma_j \text{D}(\text{TB}^c)_{t-j} \sum_{j=1}^k \delta_j \nabla \tilde{x}_{t-j}^i \quad i = A, C$$

- the “unmodified” ADF regression

$$\tilde{x}_t^B = \tilde{\phi}^B \tilde{x}_{t-1}^B + \sum_{j=1}^k \delta_j \nabla \tilde{x}_{t-j}^B + a_t$$

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. 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:

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

6.3 Unit Roots Tests When the Break Date Is Unknown

Two data-dependent methods for choosing \hat{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.

6.4 Robust Tests for a Breaking Trend

the weight function is defined as

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

- the break date is known to be at T_b^c with break fraction τ^c :

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

- When τ^c is unknown but is assumed to lie between $0 < \tau_{min}, \tau_{max} < 1$:

$$t_\lambda = \lambda(S_0(\hat{\tau}), S_1(\tilde{\tau})) \times |t_0(\hat{\tau})| + m_\xi(1 - \lambda(S_0(\hat{\tau}), S_1(\tilde{\tau}))) \times |t_1(\tilde{\tau})|$$

6.5 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

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

while $I(0)$ errors

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

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

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

6.6 Nonlinear Trends

- **logistic smooth transition** (LSTR)

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

- **exponential smooth transition** (ESTR)

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

Three alternative smooth transition trend models may then be specified as

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

HXL begin by considering robust tests of m versus at most $m - 1$ frequencies, that is, tests of:

$$H_0^*: \gamma_{1m} = \gamma_{2m} = 0; \quad \gamma_{1f}, \gamma_{2f}, f = 1, \dots, m - 1 \text{ unrestricted}$$

against

$$H_1^*: \text{at least one of } \gamma_{1m}, \gamma_{2m} \neq 0$$

The hypothesis H_0^* can be tested by the same approach using the Wald statistic:

$$W_{m-1}^m = \frac{\text{RSS}_R - \text{RSS}_U}{\text{RSS}_U/T}$$

where RSS_R now denotes the residual sum of squares from OLS estimation of (6.20) with n replaced by $m - 1$ and RSS_U is the residual sum of squares from (6.20) with n replaced by m . The modified statistic is then:

$$MW_{m-1}^m = T^{-1}W_{m-1}^m \exp\left(-\frac{b_\xi}{|DF|}\right)$$

Chapter 7 An Introduction to Forecasting With Univariate Models

7.1 Forecasting With Autoregressive Integrated-Moving Average

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. The setup is, thus, analogous to the classic statistical problem of estimating an unknown parameter of a model and providing a confidence interval for that parameter.

To formalize the forecasting problem, suppose we have a realization $(x_{1-d}, x_{2-d}, \dots, x_T)$ 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

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

then (7.1) becomes, for time $T + h$,

$$\alpha(B)x_{T+h} = \theta_0 + \theta(B)a_{T+h}$$

that is,

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

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

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

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} , and (2) replace future expectations ($j > 0$) by forecast values, $f_{T,j}$ and 0

.

The h -step ahead forecast error for origin T , defined in § 7.2, may be expressed as

$$e_{T,h} = x_{T+h} - f_{T,h} = a_{T+h} + \psi_1 a_{T+h-1} + \dots + \psi_{h-1} a_{T+1}$$

where $\psi_1, \dots, \psi_{h-1}$ are the first $h - 1$ ψ -weights in $\psi(B) = \alpha^{-1}(B)\theta(B)$. The forecast error is, therefore, a linear combination of the unobservable future shocks entering the system after time T , although the one-step ahead forecast error is simply

$$e_{T,1} = x_{T,1} - f_{T,1} = a_{T+1}.$$

Thus, for a MMSE forecast, the one-step ahead forecast errors must be uncorrelated. However, h -step ahead forecasts made at different origins will not be uncorrelated, and neither will forecasts for different lead times made at the same origin.

7.2 (ARIMA) Models Forecasting a Trend Stationary Process

Let us now 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 (7.2) is given by:

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

Since ε_t is, by assumption, stationary, we know that $g_{T,h} \rightarrow 0$ 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. As a consequence, the forecast error variance is bounded by the sample variance of ε_t , this being in contrast to the error variances of the ARIMA($p, 2, q$) process and the ARIMA($p, 1, q$) with an intercept included, which, from §§7.4 – 7.5, also have forecasts that lie on a linear trend, but have unbounded error variances. In the simplest case in which ε_t is white noise, all forecasts of a TS process have the same error variance, σ^2 .

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:

$$x_t = z_t + u_t$$

unobserved component (UC) models:

$$\nabla z_t = \mu + \gamma(B)v_t$$

and

$$u_t = \lambda(B)a_t$$

where v_t and a_t are independent white noise sequences with finite variances σ_v^2 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:

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

The assumption that the trend component, z_t , follows a random walk is not as restrictive as it may at first seem. Consider the Wold decomposition for ∇x_t :

$$\begin{aligned}\nabla x_t &= \mu + \psi(1)e_t + \nabla \tilde{\psi}(B)e_t \\ \nabla z_t &= \mu + \left(\sum_{j=0}^{\infty} \psi_j \right) e_t = \mu + \psi(1)e_t\end{aligned}$$

and

$$u_t = - \left(\sum_{j=1}^{\infty} \psi_j \right) e_t - \left(\sum_{j=2}^{\infty} \psi_j \right) e_{t-1} - \left(\sum_{j=3}^{\infty} \psi_j \right) e_{t-2} - \dots = \tilde{\psi}(B)e_t$$

Since e_t is white noise, the trend component is, therefore, a random walk with rate of drift equal to μ and an innovation equal to $\psi(1)e_t$, which is therefore proportional to that of the original series. The noise component is clearly stationary, but since it is driven by the same innovation as the trend component, z_t and u_t must be **perfectly correlated**, in direct contrast to the Muth decomposition that assumes that they are independent.

Following Newbold (1990), a straightforward way of estimating the Beveridge-Nelson components is to approximate the Wold decomposition (8.8) by an ARIMA($p, 1, q$) process by setting $\psi(B) = \theta(B)/\phi(B)$:

$$\begin{aligned}\nabla x_t &= \mu + \frac{\theta(B)}{\phi(B)} e_t = \mu + \frac{(1 - \theta_1 B - \dots - \theta_q B^q)}{(1 - \phi_1 B - \dots - \phi_p B^p)} e_t \\ \nabla z_t &= \mu + \psi(1)e_t = \mu + \frac{\theta(1)}{\phi(1)} e_t = \mu + \frac{(1 - \theta_1 - \dots - \theta_q)}{(1 - \phi_1 - \dots - \phi_p)} e_t\end{aligned}$$

8.2 Signal Extraction

Given a UC model of the form of (8.1) 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. A MMSE estimate of z_t , is an estimate \hat{z}_t which minimizes $E(\zeta_t^2)$, where $\zeta_t = z_t - \hat{z}_t$ is the estimation error (cf. §7.2). From, for example, Pierce (1979), given an infinite sample of observations, denoted $\{x_t, -\infty \leq t \leq \infty\}$, such an estimator is:

$$\hat{z}_t = \nu_z(B)x_t = \sum_{j=-\infty}^{\infty} \nu_{zj} x_{t-j}$$

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

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

in which case the noise component can be estimated as:

$$\hat{u}_t = x_t - \hat{z}_t = (1 - \nu_z(B))x_t = \nu_u(B)x_t$$

The estimation error can be written as:

$$\zeta_t = z_t - \hat{z}_t = \nu_z(B)z_t - \nu_u(B)u_t$$

and Pierce (1979) shows that ζ_t will be stationary if z_t and u_t are generated by processes of the form of (8.4). In fact, ζ_t will follow the process:

$$\zeta_t = \theta_\zeta(B)\xi_t$$

where

$$\theta_\zeta = \frac{\gamma(B)\lambda(B)}{\theta(B)} \quad \sigma_\xi^2 = \frac{\sigma_a^2\sigma_v^2}{\sigma_e^2}$$

and $\xi_t \sim \text{WN}\left(0, \sigma_\xi^2\right)$.

It can be shown that setting $\Theta = -1$ minimizes the variance of both z_t and u_t , which is known as the canonical decomposition of x_t . Choosing this value implies that $\gamma(B) = 1 + B$, and we have:

$$\hat{z}_t = \frac{\sigma_v^2(1+B)(1+B^{-1})}{\sigma_e^2(1-\theta B)(1-\theta B^{-1})}$$

and

$$(1-\theta B)\zeta_t = (1+B)\xi_t.$$

the **semi-infinite** sample $\{x_s | s \leq t-m\}$: Pierce (1979) shows that, in this case, an estimate of z_t is given by:

$$\hat{z}_t^{(m)} = \nu_z^{(m)}(B)x_t$$

where

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

8.3 Filters

The UC model (8.5) is also related to the **Hodrick-Prescott trend filter** (Hodrick and Prescott, 1997), which is a popular method of detrending economic time series. 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:

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

so that the Hodrick-Prescott (H – P) trend estimator is

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

The MMSE trend estimator can be written as:

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

In filtering terminology the H – P filter (8.18) is a **low-pass** filter. A linear filter of the observed series x_t may be defined as the two-sided weighted moving average:

$$y_t = \sum_{j=-n}^n a_j x_{t-j} = (a_{-n} B^{-n} + a_{-n+1} B^{-n+1} + \cdots + a_0 + \cdots + a_n B^n) x_t = a(B) x_t$$

The frequency response function of the filter is defined as $a(\omega) = \sum_j e^{-i\omega j}$ for a frequency $0 \leq \omega \leq 2\pi$. The **power transfer function** is then defined as:

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

and 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. 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 ω -frequency component of x_t is displaced in time. If the filter is indeed symmetric then $a(\omega) = a(-\omega)$, so that $a(\omega) = |a(\omega)|$ and $\theta(\omega) = 0$, known as phase neutrality.

With these concepts, an "ideal" low-pass filter has the frequency response function:

$$a_L(\omega) = \begin{cases} 1 & \text{if } \omega < \omega_c \\ 0 & \text{if } \omega > \omega_c \end{cases}$$

or

$$a_L(B) = \frac{\omega_c}{\pi} + \sum_{j=1}^{\infty} \frac{\sin \omega_c j}{\pi j} (B^j + B^{-j})$$

Chapter 9 Seasonality and Exponential Smoothing

9.1 Seasonal Patterns in Time Series

The presence of seasonality is often immediately apparent from a plot of the series, but it will also manifest itself in the sample autocorrelation function (SACF) of the appropriately differenced data.

9.2 Modeling Deterministic Seasonality

A simple model is to use a "seasonal mean" model in which there is a different mean for each season, that is, the model for x_t is:

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

The regression model thus assumes that the seasonal pattern is **deterministic**, in the sense that the seasonal means α_i , $i = 1, \dots, m$, remain constant through time.¹

9.3 Modeling Stochastic Seasonality

The slow decline of these seasonal autocorrelations is indicative of ***seasonal nonstationarity*** and, analogous to the analysis of "nonseasonal nonstationarity," this may be dealt with by ***seasonal differencing***, that is, by using the $\nabla_4 = 1 - B^4$ operator in conjunction with the usual ∇ operator.

In general, if we have a seasonal period of m then the seasonal differencing operator may be denoted as ∇_m . The nonseasonal and seasonal differencing operators may then be applied d and D times, respectively, so that a seasonal ARIMA model may take the general form

$$\nabla^d \nabla_m^D \phi(B) x_t = \theta(B) a_t$$

the general multiplicative seasonal model:

$$\phi_p(B) \Phi_P(B^m) \nabla^d \nabla_m^D x_t = \theta_q(B) \Theta_Q(B^m) a_t$$

The subscripts p, P, q, Q have been added for clarity so that the orders of the various polynomials may be emphasized and the ARIMA process in (9.5) is said to be of order $(p, d, q)(P, D, Q)_m$.

Because the general multiplicative model is rather complicated, explicit expressions for its ACF and PACF are difficult to provide. This led Box and Jenkins to consider a particularly simple case, in which an ARIMA $(0, 1, 1)$ is used to link the x_t s one ***year*** apart:

$$\nabla_m x_t = (1 - \Theta B^m) \alpha_t$$

and a similar model is used to link α_t s one ***observation*** apart:

$$\nabla \alpha_t = (1 - \theta B) a_t$$

where, in general, θ and Θ will have different values. On combining the two equations we obtain the multiplicative ARIMA $(0, 1, 1)(0, 1, 1)_m$ model:

$$\nabla \nabla_m x_t = (1 - \theta B)(1 - \Theta B^m) \alpha_t$$

the variances for the estimated sample autocorrelations at lags higher than $m + 1$ are given by:

$$V(r_k) = T^{-1} (1 + 2(r_1^2 + r_{m-1}^2 + r_m^2 + r_{m+1}^2)) \quad k > m + 1$$

Using this result in conjunction with the known form of the ACF will enable the ARIMA $(0, 1, 1)(0, 1, 1)_m$ model to be identified.

Forecasts for the ARIMA $(0, 1, 1)(0, 1, 1)_m$ model may be computed using the approach outlined in § 7.2, so that:

$$f_{T,h} = E(x_{T+h-1} + x_{T+h-m} - x_{T+h-m-1} + a_{T+h} - \theta a_{T+h-1} - \Theta a_{T+h-m} + \theta \Theta a_{T+h-m-1} \mid x_T, x_{T-1}, \dots)$$

9.4 Mixed Seasonal Models

The deterministic and stochastic seasonal models, (9.1) and (9.5), may be combined to form, on setting $d = D = 1$ for both simplicity and because these are the settings that are typically found,

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

9.5 Seasonal Adjustment

In § 2.16 we introduced a decomposition of an observed time series into trend, seasonal, and irregular (or noise) components, focusing attention on estimating the seasonal component and then eliminating it to provide a seasonally adjusted series. Extending the notation introduced in (8.1), this implicit UC decomposition can be written as

$$x_t = z_t + s_t + u_t$$

where the additional seasonal component s_t is assumed to be independent of both z_t and u_t . On obtaining an estimate of the seasonal component, \hat{s}_t , the seasonally adjusted series can then be defined as $x_t^a = x_t - \hat{s}_t$.

9.6 Exponential Smoothing

Returning to the two-component UC model, (8.1), 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 :

$$(1 - (1 - \alpha)B)z_t = \alpha x_t$$

or,

$$z_t = \alpha x_t + (1 - \alpha)z_{t-1}$$

which is the basic algorithm of simple (or single) **exponential smoothing**

Simple exponential smoothing is, therefore, a suitable forecasting procedure for a series in which a trend is absent. To capture a linear trend, the approach may be generalized by extending (9.11) to include a trend component,

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

where the error correction is now $e_t = x_t - z_{t-1} - \tau_{t-1}$, and defining a second updating equation for the trend τ_t :

$$\begin{aligned} \tau_t &= \beta(z_t - z_{t-1}) + (1 - \beta)\tau_{t-1} \\ &= \tau_{t-1} + \alpha\beta e_t \end{aligned}$$

This pair of updating equations are together known as the Holt-Winters model.⁸ Forecasts are given by:

$$f_{T,h} = z_T + \tau_T h$$

A related approach is that of **double exponential smoothing**, which is defined by the pair of recursions:

$$\begin{aligned} z_t &= \gamma x_t + (1 - \gamma)z_{t-1} \\ \tau_t &= \gamma(z_t - z_{t-1}) + (1 - \gamma)\tau_{t-1} \end{aligned}$$

so that only a single smoothing parameter is used.

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 x_t is generated by the **product process**:

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

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

$$V(x_t | \sigma_t) = E((x_t - \mu)^2 | \sigma_t) = \sigma_t^2 E(U_t^2) = \sigma_t^2$$

σ_t^2 is, thus, the **conditional variance** and σ_t the **conditional standard deviation** of x_t .

10.2 Autoregressive Conditional

Up until this point we have said nothing about how the conditional variances σ_t^2 might be generated. We now consider the case where they are a function of past values of x_t :

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

A simple example is:

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

where α_0 and α_1 are both positive to ensure that $\sigma_t^2 > 0$.

A natural extension is to the ARCH(q) process, where (10.2) is replaced by:

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

where $\alpha_i \geq 0, 0 \leq i \leq q$.

To obtain more flexibility, a further extension, to the generalized ARCH (GARCH) process, was introduced by Bollerslev (1986). The GARCH(p, q) process has the conditional variance function:

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

where $p > 0$ and $\beta_i \geq 0, i \leq 1 \leq p$.

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

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

10.3 Heteroskedastic Processes

An asymmetric response to shocks is made explicit in the **exponential** GARCH (EGARCH) model of Nelson (1991):

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

where

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

A model which nests (10.4)-(10.6) is the nonlinear ARCH model of Higgins and Bera (1992), a general form of which is:

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

while an alternative is the threshold ARCH process:

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

where

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

$\mathbf{1}(\cdot)$ being the indicator function

10.4 Testing for the Presence of

Suppose we have the ARMA(P, Q)-GARCH (p, q) model of § 10.11:

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

Forecasts of x_{T+h} can be obtained from the "mean equation" (10.7) in the manner outlined in §§ 7.1-7.4. When calculating forecast error variances, however, it can no longer be assumed that the error variance itself is constant. Thus, (7.4) must be amended to:

$$V(e_{t,h}) = \sigma_{T+h}^2 + \psi_1^2 \sigma_{T+h-1}^2 + \cdots + \psi_{h-1}^2 \sigma_{T+1}^2$$

with the σ_{T+h}^2 being obtained recursively from (10.8).

Chapter 11 Nonlinear Stochastic Processes

11.1 Martingales, Random Walks, and Nonlinearity

A martingale may be defined as a stochastic process x_t having the following properties:²

1. $E(|x_t|) < \infty$ for each t ;
2. $E(x_t | x_s, x_{s-1}, \dots) = x_s$.

Written as

$$E(x_t - x_s | x_s, x_{s-1}, \dots) = 0, \quad s < t,$$

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:

$$E(x_t - x_s | x_s, x_{s-1}, \dots) \geq 0, \quad s < t,$$

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

The martingale given by (11.1) can be written equivalently as:

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

where a_t is known as the martingale increment or ***martingale difference***. When written in this form, x_t looks superficially identical to a random walk.

The possibility of this form of dependence leads naturally to the consideration of ***nonlinear*** stochastic processes capable of modeling such behavior. As an illustration, suppose that x_t is generated by the process $\nabla x_t = \eta_t$, with η_t being defined as:

$$\eta_t = a_t + \beta a_{t-1} a_{t-2}$$

11.2 Nonlinear Stochastic Models

A stochastic process can then be considered nonlinear if it does not satisfy the assumptions underlying the decomposition, for example, if the representation is:

$$x_t - \mu = f(a_t, a_{t-1}, a_{t-2}, \dots)$$

where $f(\cdot)$ is some arbitrary nonlinear function.

Polynomial functions of lagged x_t can also be used (Jones, 1978), while another simple way of introducing nonlinearity is to allow x_t to respond in a different manner to innovations depending on their sign. For example, Wecker (1981) introduced the asymmetric moving average process, whose first-order form is:

$$x_t = a_t + \theta a_{t-1} + \psi \mathbf{1}(a_{t-1} > 0) a_{t-1}$$

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

11.3 Bilinear Models

An important class of nonlinear model is the bilinear, which takes the general form

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

Here $\varepsilon_t \sim SWN(0, \sigma_\varepsilon^2)$.

Granger and Andersen (1978) have analyzed the properties of several simple bilinear forms, characterized as:

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

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

Charemza et al. (2005) discuss nonstationary generalizations of bilinear models that allow for unit roots. For example, they consider the following simple model:

$$x_t = (a + b\varepsilon_{t-1})x_{t-1} + \varepsilon_t$$

Maravall (1983) considers an alternative form of bilinearity in which x_t is given by the ARMA process,

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

but where the uncorrelated innovation sequence is bilinear in a_t and the strict white noise sequence ε_t :

$$a_t = \varepsilon_t + \sum_{i=1}^R \sum_{j=1}^S \gamma_{ij} a_{t-i} \varepsilon_{t-j}$$

This may be interpreted as a bilinear model "forecasting white noise."

11.4 Threshold and Smooth Transition Autoregressions

A popular class of nonlinear model is the ***self-exciting*** threshold autoregressive (SETAR) process, which allows for asymmetry by defining a set of piecewise autoregressive models whose switch points, or "thresholds," are generally unknown (see Tong and Lim, 1980; Tong, 1990; Teräsvirta, 2006):

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

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 $a_{j,t} \sim WN(0, \sigma_j^2)$, $j = 1, \dots, r$, so that the error variance is allowed to alter across the r "regimes." A popular version is the two-regime SETAR ($2 : p, d$) model:

$$\begin{aligned} x_t = & (\phi_{1,1} x_{t-1} + \cdots + \phi_{1,p} x_{t-p} + a_{1,t}) \mathbf{1}(x_{t-d} \leq c_1) \\ & + (\phi_{2,1} x_{t-1} + \cdots + \phi_{2,p} x_{t-p} + a_{2,t}) (1 - \mathbf{1}(x_{t-d} \leq c_1)) \end{aligned}$$

The SETAR formulation requires that the shift from one regime to another is immediate. Allowing the shift to be smooth is accomplished by defining the ***exponential autoregressive*** (EAR) process:

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

where the transition function

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

is symmetric around zero, where it takes the value unity, and as $|x_{t-d}| \rightarrow \infty$ so $G(\gamma, x_{t-d}) \rightarrow 0$.

This model, originally proposed by Haggan and Ozaki (1981), was subsequently extended by Teräsvirta (1994) to allow for asymmetry in the transition function by including a location parameter c :

$$G_E(\gamma, c, x_{t-d}) = \exp(-\gamma(x_{t-d} - c)^2), \quad \gamma > 0$$

This is known as the ***exponential smooth transition*** AR [ESTAR (p, d)] model, while an alternative specification of the transition function produces the ***logistic*** STAR (LSTAR) model:

$$G_L(\gamma, c, x_{t-d}) = (1 + \exp(-\gamma(x_{t-d} - c)))^{-1}, \quad \gamma > 0$$

Note that when $\gamma = 0$, $G_L(\gamma, c, x_{t-d}) = 0.5$ and (11.8) reduces to a linear AR model, while if $\gamma \rightarrow \infty$ the LSTAR model approaches the SETAR, albeit with $\sigma_1 = \sigma_2$. If t replaces x_{t-d} in (11.9) then the resulting model is referred to as a ***time-varying autoregression***, which enables testing for the null of parameter constancy in linear AR models, with smoothly changing parameters forming the alternative to the null: see Lin and Teräsvirta (1994).

11.5 Markov-Switching Models

Yet another way of introducing asymmetry is to consider "regime switching" models. Hamilton (1989, 1990), Engle and Hamilton (1990), and Lam (1990) all propose variants of a switching-regime Markov model, which can be regarded as a nonlinear extension of an ARMA process that can accommodate complicated dynamics, such as asymmetry and conditional heteroskedasticity. 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

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

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, so that, unlike the conventional UC specification, u_t can be nonstationary.

ML estimates of the model are obtained by using to write the noise component as

$$u_t = x_t - \alpha_0 - \alpha_1 S_t$$

the innovations $\varepsilon_t = \phi(B)u_t$ can be expressed as

$$\varepsilon_t = \phi(B)(x_t - \alpha_0 - \alpha_1 S_t)$$

11.6 Neural Networks

Neural networks (NNs) refer to a broad class of nonparametric models which have gained a good deal of popularity in recent years across a wide range of disciplines, including computer science, psychology, biology, linguistics, and pattern recognition (for a textbook treatment, see, for example, Haykin, 1999).

A univariate autoregressive MLP model with a single hidden layer can be represented as:

$$x_t = \sum_{i=1}^p \phi_i x_{t-i} + \sum_{j=1}^q \beta_j G \left(\sum_{i=1}^p \varphi_i x_{t-i} \right) + \varepsilon_t$$

11.7 Nonlinear Dynamics and Chaos

Research in the general area of nonlinear dynamics is concerned with the behavior of deterministic and stochastic nonlinear systems. This diversity of meanings is mainly a consequence of there being no formal and complete mathematical definition of a chaotic system (see, for example, Berliner, 1992).

11.8 Testing for Nonlinearity

Ramsey's original Regression Error Specification Test (RESET) is 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 $H_0 : \delta_j = 0, j = 2, \dots, h$.

Once evidence in favor of nonlinearity has been found, Hsieh (1989) has developed a test that can shed light on the type of nonlinearity present. More specifically, the test attempts to discriminate between two types of nonlinearity: "additive" and "multiplicative." In the former, nonlinearity enters solely through the conditional mean of the process:

$$e_t = g(x_{t-1}, \dots, x_{t-k}, e_{t-1}, \dots, e_{t-k}) + u_t$$

where $g(\cdot)$ is an arbitrary nonlinear function. This suggests that a bilinear, threshold, or smooth transition model may be appropriate. Multiplicative nonlinearity manifests itself through the conditional variance, thus pointing toward an ARCH-type model:

$$e_t = g(x_{t-1}, \dots, x_{t-k}, e_{t-1}, \dots, e_{t-k})u_t$$

The test exploits the fact that, unlike additive dependence, multiplicative dependence implies that

$$E(e_t | x_{t-1}, \dots, x_{t-k}, e_{t-1}, \dots, e_{t-k}) = 0$$

Ramsey and Rothman (1996) have proposed the TR test statistic, estimated for various lags k as:

$$TR(k) = \hat{B}_{2,1}(k) - \hat{B}_{1,2}(k)$$

where $\hat{B}_{2,1}(k)$ and $\hat{B}_{1,2}(k)$ are estimators of the **bicovariances** $E(x_t^2 x_{t-k})$ and $E(x_t x_{t-k}^2)$, respectively. These can be estimated using the residuals from a linear fit as:

$$\hat{B}_{i,j}(k) = (T - k)^{-1} \sum_{t=k+1}^T e_t^i e_{t-k}^j \quad i, j = 1, 2$$

A nonparametric test that has created considerable interest is the BDS (standing for Brock-Dechert-Scheinkman) statistic, based on the concept of the **correlation integral**: see, for example, Brock (1986), Brock et al. (1991), and Dechert (1996). The test is based on the idea that the evolution of the next values of any two blocks of observations, which are close in some metric, should also be close in the same metric. For an observed series x_t , the correlation integral $C_N(\ell, T)$ is defined as:

$$C_N(\ell, T) = \frac{2}{T_N(T_N - 1)} \sum_{t < s} \mathbf{I}_t(x_t^N, x_s^N)$$

where

$$x_t^N = (x_t, x_{t+1}, \dots, x_{t+N-1})$$

and

$$x_s^N = (x_s, x_{s+1}, \dots, x_{s+N-1})$$

are called " N -histories," $\mathbf{I}_t(x_t^N, x_s^N)$ equals one if $\|x_t^N - x_s^N\| < \ell$ and zero otherwise, $\|\cdot\|$ being the sup-norm, and $T_N = T - N + 1$.

The correlation integral is an estimate of the probability that any two N histories, x_t^N and x_s^N , are within ℓ of each other. If the x_t s are strict white noise, then

$$C_N(\ell, T) \rightarrow C_1(\ell, T)^N, \text{ as } T \rightarrow \infty$$

and

$$w_N(\ell, T) = \frac{\sqrt{T} (C_N(\ell, T) - C_1(\ell, T)^N)}{\sigma_N(\ell, T)}$$

has a standard normal limiting distribution, where the expression for the variance $\sigma_N^2(\ell, T)$ may be found in.

Tests are also available for specific nonlinear alternatives. Tests against ARCH and bilinear alternatives have already been discussed in § 11.15-11.16 and there is also a fully developed testing procedure against STAR models. From Teräsvirta (1994), an LM-type test statistic for the null of linearity against an LSTAR alternative can be constructed from the auxiliary regression,

$$e_t = \sum_{i=1}^p \varphi_i x_{t-i} + \sum_{i=1}^p \delta_{1j} x_{t-i} x_{t-d} + \sum_{i=1}^p \delta_{2j} x_{t-i} x_{t-d}^2 + \sum_{i=1}^p \delta_{3j} x_{t-i} x_{t-d}^3 + v_t$$

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

It should be emphasized, however, that all these tests are designed to distinguish between linear and nonlinear **stochastic** dynamics.

11.9 Forecasting With Nonlinear Models

One-step ahead forecasting from nonlinear models is straightforward, but multistep forecasting may be complicated. If we have the simple nonlinear model, $x_t = g(x_{t-1}) + \varepsilon_t$ then the one-step ahead forecast is straightforwardly given by $f_{T,1} = g(x_T)$. However, the two-step ahead forecast is

$$f_{T,2} = E(x_{T+2} | x_T) = E(g(g(x_{T+1}) + \varepsilon_{T+1})) = \int_{\varepsilon} g(g(x_{T+1}) + \varepsilon_{T+1}) dF(\varepsilon)$$

where $dF(\varepsilon)$ is the cumulative distribution function of ε_t .

Chapter 12 Transfer Functions and Autoregressive Distributed Lag Modeling

12.1 Transfer Function-Noise Models

We shall analyze several **multivariate** models over the next chapters, but our development begins with the simplest. This is the **single-input transfer function-noise** model, in which an **endogenous**, or output, variable y_t is related to a single input, or **exogenous**, variable x_t through the dynamic model ¹

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

where the lag polynomial $v(B) = v_0 + v_1B + v_2B^2 + \dots$ 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**.

More precisely, $v(B)$ 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

$$\omega(B) = \omega_0 - \omega_1B - \dots - \omega_sB^s$$

and

$$\delta(B) = 1 - \delta_1 B - \cdots - \delta_r B^r$$

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

The noise process may be assumed to follow an ARMA (p, q) model:

$$n_t = \frac{\theta(B)}{\phi(B)} a_t$$

so that the combined transfer function-noise model can be written as

$$y_t = \frac{\omega(B)}{\delta(B)} x_{t-b} + \frac{\theta(B)}{\phi(B)} a_t$$

If identifying a univariate ARMA model is often considered to be an "art form," then identifying a transfer function in this way is even more so and, if there are multiple inputs, can become increasingly difficult, for the model is now:

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

where

$$\omega_j(B) = \omega_{j,0} - \omega_{j,1} B - \cdots - \omega_{j,s_j} B^{s_j}$$

and

$$\delta_j(B) = 1 - \delta_{j,1} B - \cdots - \delta_{j,r_j} B^{r_j}$$

12.2 Autoregressive Distributed Lag Models

This restricted form is known as the autoregressive distributed lag, or ARDL, model and is obtained by placing the following restrictions on (12.4):

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

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

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

or

$$y_t = \beta_0 + \sum_{i=1}^p \phi_i y_{t-i} + \sum_{j=1}^M \beta_j(1)x_{j,t} + \sum_{j=1}^M \tilde{\beta}_j(B)\nabla x_{j,t} + a_t$$

Solving for y_t then yields

$$y_t = \theta_0 + \sum_{j=1}^M \theta_j x_{j,t} + \sum_{j=1}^M \tilde{\theta}_j(B)\nabla x_{j,t} + \varepsilon_t$$

in which

$$\begin{aligned} \theta_0 &= \phi^{-1}(1)\beta_0 \\ \theta_j &= \phi^{-1}(1)\beta_j(1) \\ &= \phi^{-1}(B) \left(\tilde{\beta}_j(B) - \tilde{\phi}(B)\phi^{-1}(1)\beta_j(1) \right) \quad j = 1, \dots, M \\ \varepsilon_t &= \phi^{-1}(B)a_t \end{aligned}$$

where

$$\tilde{\phi}(B) = \nabla^{-1}(\phi(B) - \phi(1))$$

has been used.

This is known as the ARDL (p, s_1, \dots, s_M) model.

The third representation separates out the long-run relationships between the output and the inputs from short-run effects, but is not amenable to direct estimation. The estimate of the long-run relationship between y and x_j may be obtained from the second representation and is given by

$$\hat{\theta}_j = \frac{\hat{\beta}_j(1)}{1 - \sum_{i=1}^p \hat{\phi}_i} = \frac{\hat{\beta}_{j,0} + \dots + \hat{\beta}_{j,s_j}}{1 - \sum_{i=1}^p \hat{\phi}_i}$$

Chapter 13 Vector Autoregressions and Granger Causality

13.1 Multivariate Dynamic Regression Models

In a natural extension to the ARDL model of the previous chapter, suppose that there are now two endogenous variables, $y_{1,t}$ and $y_{2,t}$, that may both be related to an exogenous variable x_t and its lags as well as to lags of each other. In the simplest case, such a model would be:

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

The "system" contained in Eq. (13.1) is known as a multivariate dynamic **regression**, a model treated in some detail in Spanos (1986, Chapter 24).

the general form of the multivariate dynamic regression model:

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

where there is a maximum of p lags on the endogenous variables and a maximum of q lags on the exogenous variables. Here $\mathbf{c}' = (c_1, c_2, \dots, c_n)$ is a $1 \times n$ vector of constants and $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_p$ and $\mathbf{B}_0, \mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_q$ are sets of $n \times n$ and $n \times k$ matrices of regression coefficients, respectively, such that

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

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

13.2 Vector Autoregressions

Suppose 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 it is now simply a p th order autoregression in the vector \mathbf{y}_t it is known as a ***vector autoregression*** (VAR(p)) of dimension n and, again, can be estimated by multivariate least squares.

13.3 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. The presence of such dynamic relationships is known as ***Granger (-Sims) causality***. The variable y_s **does not Granger-cause** the variable y_r if $a_{rs,i} = 0$ for all $i = 1, 2, \dots, p$. If, on the other hand, there is at least one $a_{rs,i} \neq 0$ then y_s is said to **Granger-cause** y_r because if that is the case then past values of y_s are useful in forecasting the current value of y_r : Granger causality is, thus, a criterion of "forecastability." If y_r also Grangercauses y_s , the pair of variables are said to exhibit **feedback**.

Within a VAR(p), Granger causality running from y_s to y_r , which may be depicted as $y_s \rightarrow y_r$, can be evaluated by setting up the null hypothesis of non-Granger causality (y_s does not $\rightarrow y_r$), $H_0 : a_{rs,1} = \dots = a_{rs,p} = 0$, and testing this with a Wald statistic; a multivariate extension of the standard F -statistic for testing a set of zero restrictions in a conventional regression model: see, for example, Mills (2014, §13.3).

13.4 Determining the Lag Order of a Vector Autoregression

Consider VAR(p) with error covariance matrix $\Omega_p = E(\mathbf{u}_t \mathbf{u}_t')$, 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:

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

where $\hat{\mathbf{U}}_p = (\hat{\mathbf{u}}'_{p,1}, \dots, \hat{\mathbf{u}}'_{p,n})'$ is the matrix of residuals obtained by OLS estimation of the VAR(p), $\hat{\mathbf{u}}_{p,r} = (\hat{u}_{r,p+1}, \dots, \hat{u}_{r,T})'$ being the residual vector from the r th 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

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

Thus, if $LR(p, m)$ exceeds the α critical value of the χ^2 distribution with $n^2(p - m)$ degrees of freedom, then the hypothesis that the VAR order is m is rejected at the α level of significance in favor of the higher order p .

multivariate AIC and BIC criteria:

$$\begin{aligned} MAIC(p) &= \log |\hat{\Omega}_p| + (2 + n^2 p) T^{-1} \\ MBIC(p) &= \log |\hat{\Omega}_p| + n^2 p T^{-1} \ln T \quad p = 0, 1, \dots, p_{\max} \end{aligned}$$

13.5 Variance Decompositions and Innovation Accounting

This has led to the development of several techniques for examining the "information content" of a VAR that are based on the **vector moving average representation** (VMA) of \mathbf{y}_t . Suppose that the VAR is written in lag operator form as

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

where

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

is a matrix polynomial in B . Analogous to the univariate case (recall § 3.8 – 3.9), the (infinite order) VMA representation is

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

where

$$\Psi_i = \sum_{j=1}^i \mathbf{A}_j \Psi_{i-j} \quad \Psi_0 = \mathbf{I}_n \quad \Psi_i = \mathbf{0} \quad i < 0$$

The Ψ_i matrices can be interpreted as the **dynamic multipliers** of the system, since they represent the model's response to a unit shock in each of the variables. The response of y_r to a unit shock in y_s (produced by $u_{s,t}$ taking the value unity rather than its expected value of zero) is, therefore, given by the **impulse response function**, which is the sequence $\psi_{rs,1}, \psi_{rs,2}, \dots$, where $\psi_{rs,i}$ is the rs th element of the matrix Ψ_i .

The impulse response function of y_r to a y_s shock is then given by the sequence $\psi_{rs,0}^O, \psi_{rs,1}^O, \psi_{rs,2}^O, \dots$, where each impulse response can be written compactly as:

$$\psi_{rs,i}^O = \mathbf{e}'_r \Psi_i \mathbf{S} \mathbf{e}_s$$

Here \mathbf{e}_s is the $n \times 1$ selection vector containing unity as the s th element and zero elsewhere. This sequence is known as the **orthogonalized impulse response function**. The (**accumulated long-run response**) is then:

$$\psi_{rs}^O(\infty) = \sum_{i=0}^{\infty} \mathbf{e}'_r \Psi_i \mathbf{S} \mathbf{e}_s$$

The entire set of long-run responses may then be gathered together in the matrix

$$\Psi^o(\infty) = \sum_{i=0}^{\infty} \Psi_i \mathbf{S} = \Psi(1) \mathbf{S}$$

Apart from comparing the impulse responses and variance decompositions for alternative orderings of the variables, one solution to this problem is to use Pesaran and Shin's (1997) generalized impulse responses, defined by replacing \mathbf{S} in with $\sigma_r^{-1} \boldsymbol{\Omega}_p$:

$$\psi_{rs,i}^G = \sigma_r^{-1} \mathbf{e}'_r \Psi_i \boldsymbol{\Omega}_p \mathbf{e}_s$$

13.6 Structural Vector Autoregressions

The Cholesky decomposition of § 13.12 can be written as $\mathbf{u}_t = \mathbf{S}\mathbf{v}_t$ with $\mathbf{S}\mathbf{S}' = \boldsymbol{\Omega}_p$ and $E(\mathbf{v}_t\mathbf{v}'_t) = \mathbf{I}_n$. 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} .

An alternative form of restriction is also possible. The long-run impulse response may be written, on generalizing (13.7), as

$$\psi_{rs}(\infty) = \sum_{i=0}^{\infty} \mathbf{e}'_r \boldsymbol{\Psi}_i \mathbf{A}^{-1} \mathbf{B} \mathbf{e}_s$$

or, in matrix form,

$$\boldsymbol{\Psi}(\infty) = \sum_{i=0}^{\infty} \boldsymbol{\Psi}_i \mathbf{A}^{-1} \mathbf{B} = \boldsymbol{\Psi}(1) \mathbf{A}^{-1} \mathbf{B}$$

Chapter 14 Error Correction, Spurious Regressions, and Cointegration

14.1 The Error Correction Form of an Autoregressive Distributed Lag

ARDL(1, 1):

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

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

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

the **error-correction model** or ECM:

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

the ECM of general ARDL (p, s_1, \dots, s_M) model:

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

where

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

14.2 Spurious Regressions

Granger and Newbold considered the following data generation process (DGP):

$$\begin{aligned}y_t &= \phi y_{t-1} + u_t \quad u_t \sim \text{i.i.d. } (0, \sigma_u^2) \\ x_t &= \phi^* x_{t-1} + v_t \quad v_t \sim \text{i.i.d. } (0, \sigma_v^2) \\ &= 0 \text{ for all } t, s \quad E(u_t u_{t-k}) = E(v_t v_{t-k}) = 0 \text{ for all } k \neq 0\end{aligned}$$

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

More generally, if $y_t \sim I(d)$ and $x_t \sim I(d)$ then the linear combination

$$e_t = y_t - ax_t$$

will usually be $I(d)$ as well. It is possible, however, that e_t may be integrated of a lower order, say $I(d-b)$, where $b > 0$, in which case a special constraint operates on the long-run components of the two series.

Cointegration and error correction are also intimately linked. The equilibrium error (14.8), which will be $I(0)$ if y_t and x_t are both $I(1)$ and cointegrated, is exactly of the form ec_t that enters with a lag in the ECM (14.4). Consequently, if y_t and $x_{1,t}, x_{2,t}, \dots, x_{M,t}$ are all $I(1)$ and if y is cointegrated with the x 's such that

$$e_t = y_t - a_1 x_{1,t} - \cdots - a_M x_{M,t} \sim I(0)$$

then there will exist an ECM of the form of (14.4), a result that is known as **Granger's Representation Theorem**, which is proved in a more general framework by Engle and Granger (1987)

14.4 Testing for Cointegration

The conditional error correction (CEC) form of an ARDL model is, with a time trend included,

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

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

Although it is the general form of the CEC, there are several ways in which the constant and trend can enter the error correction.

14.5 Estimating Cointegrating Regressions 250

Having found that y_t cointegrates with $x_{1,t}, \dots, x_{M,t}$, the parameters in the cointegrating regression

$$y_t = \beta_0 + \beta_1 x_{1,t} + \dots + \beta_M x_{M,t} + e_t$$

OLS:

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

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

15.1 Vector Autoregressions With Integrated Variables

the n -variable VAR(p):

$$A(B)y_t = c + u_t$$

where, as in § 13.2, $E(\mathbf{u}_t) = \mathbf{0}$ and

$$E(\mathbf{u}_t \mathbf{u}_s) = \begin{cases} \boldsymbol{\Omega}_p & t = s \\ \mathbf{0} & t \neq s \end{cases}$$

an equivalent representation is

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

where

$$\Pi = \mathbf{A} - \mathbf{I}_n = -\mathbf{A}(1)$$

Consider first the case where $\mathbf{A} = \mathbf{I}_n$, so that $\Pi = \mathbf{0}$ and $\nabla \mathbf{y}_t$ follows the VAR($p-1$)

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

where \mathbf{y}_t is an $I(1)$ process and a VAR in the first differences $\nabla \mathbf{y}_t$, is the appropriate specification.

15.2 Vector Autoregressions With Cointegrated Variables

The condition $\mathbf{A} = \mathbf{I}_n$ implies that

$$|\Pi| = |\mathbf{A}_1 + \cdots + \mathbf{A}_p - \mathbf{I}_n| = 0$$

that is, that the long-run matrix is singular and must, therefore, have a rank that is less than n . Substituting $\Pi = \beta\alpha'$ into (15.2) yields

$$\nabla y_t = c + \Phi(B)\nabla y_{t-1} + \beta\alpha'y_{t-1} + u_t$$

If \mathbf{y}_t is cointegrated with cointegrating rank r , then it can be represented as the **vector error correction model** (VECM)

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

where $\mathbf{e}_t = \alpha'\mathbf{y}_t$ contains the r stationary error corrections. This is known as **Granger's Representation Theorem**.

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

VECM written as:

$$\nabla \mathbf{y}_t = c + \sum_{i=1}^{p-1} \Phi_i \nabla \mathbf{y}_{t-i} + \beta \alpha' \mathbf{y}_{t-1} + \mathbf{u}_t$$

Treat it as a generalized eigenvalue problem and solving a set of equations of the form:

$$(\lambda_i \mathbf{S}_{11} - \mathbf{S}_{10} \mathbf{S}_{00}^{-1} \mathbf{S}_{01}) \mathbf{v}_i = 0 \quad i = 1, \dots, n$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ are the set of eigenvalues and $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$ contains the set of associated eigenvectors, subject to the normalization

$$\mathbf{V} \mathbf{S}_{11} \mathbf{V} = \mathbf{I}_n$$

The ML estimate of α is then given by the eigenvectors corresponding to the r largest eigenvalues:

$$\hat{\alpha} = (v_1, v_2, \dots, v_r)$$

and the ML estimate of β is consequently calculated as $\hat{\beta} = \mathbf{S}_{01}\hat{\alpha}$, which is equivalent to the estimate of β that would be obtained by substituting $\hat{\alpha}$ into (15.7) and estimating by OLS, which also provides ML estimates of the remaining parameters in the model.

15.4 Identification of Vector Error Correction Models

the VECM as the pair of "marginal" models:

$$\begin{aligned} \nabla \mathbf{x}_t &= \mathbf{c}_1 + \sum_{i=1}^{p-1} \Phi_{1,i} \nabla \mathbf{y}_{t-i} + \beta_1 \alpha' \mathbf{y}_{t-1} + \mathbf{u}_{1,t} \\ \nabla \mathbf{z}_t &= \mathbf{c}_2 + \sum_{i=1}^{p-1} \Phi_{2,i} \nabla \mathbf{y}_{t-i} + \beta_2 \alpha' \mathbf{y}_{t-1} + \mathbf{u}_{2,t} \end{aligned}$$

where

$$\Phi_i = \begin{bmatrix} \Phi_{1,i} \\ \Phi_{2,i} \end{bmatrix} \quad i = 1, \dots, m-1 \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \quad u_t = \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}$$

are conformable partitions. The condition for z_t to be weakly exogenous for (α, β_1) is $\beta_2 = \mathbf{0}$, in which case the error correction $e_t = \alpha' y_t$ does not enter the marginal model for z_t ²²

15.5 Structural Vector Error Correction Models

Following Johansen and Juselius (1994), a "structural VECM" may be written as

$$\Gamma_0 \nabla y_t = \sum_{i=1}^{p-1} \Gamma_i \nabla y_{t-i} + \Theta \alpha y_{t-1} + v_t$$

which is related to the "reduced form" VECM

$$\nabla y_t = \sum_{i=1}^{p-1} \Phi_i \nabla y_{t-i} + \beta \alpha y_{t-1} + u_t$$

through

$$\begin{aligned} \Gamma_i &= \Gamma_0 \Phi_i \quad i = 1, \dots, p-1 \\ \Gamma_0 \beta &= \Theta \quad \nu_t = \Gamma_0 u_t \end{aligned}$$

so that

$$E(\nu_t \nu_t') = \Gamma_0 \Omega_p \Gamma_0'$$

15.6 Causality Testing in Vector Error Correction Models

Consider a "fully partitioned" form of the marginal VECM (15.10a,b):

$$\begin{aligned} \nabla x_t &= c_1 + \sum_{i=1}^{p-1} \Phi_{11,i} \nabla x_{t-i} + \sum_{i=1}^{p-1} \Phi_{12,i} \nabla z_{t-i} + \beta_1 \alpha'_1 x_{t-1} + \beta_1 \alpha'_2 z_{t-1} + u_{1,t} \\ \nabla z_t &= c_2 + \sum_{i=1}^{p-1} \Phi_{21,i} \nabla x_{t-i} + \sum_{i=1}^{p-1} \Phi_{22,i} \nabla z_{t-i} + \beta_2 \alpha'_1 x_{t-1} + \beta_2 \alpha'_2 z_{t-1} + u_{2,t} \end{aligned}$$

where now

$$\Phi_i = \begin{bmatrix} \Phi_{11,i} & \Phi_{12,i} \\ \Phi_{21,i} & \Phi_{22,i} \end{bmatrix} \quad \alpha' = [\alpha_1 \quad \alpha_2]'$$

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

$$\mathcal{H}_0 : \Phi_{12,1} = \dots = \Phi_{12,p-1} = \mathbf{0}, \quad \beta_1 \alpha'_2 = \mathbf{0}$$

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.

15.7 Impulse Response Asymptotics in Nonstationary VARs

As shown in § 13.11-13.14, the various impulse responses of the VAR are computed from the sequence of matrices

$$\Psi_i = \sum_{j=1}^i A_j \Psi_{i-j}, \quad \Psi_0 = I_n \quad \Psi_i = \mathbf{0}, \quad i < 0$$

Their computation remains exactly the same in nonstationary VARs, but if $\Pi = -\sum_{j=1}^p A_j$ is of reduced rank, the elements of Ψ_i will not die out as i increases, and this leads to some analytical complications.

15.8 Vector Error Correction Model-X Models

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

$$\nabla \mathbf{y}_t = \mathbf{c} + \mathbf{d}t + \sum_{i=1}^{p-1} \Phi_i \nabla \mathbf{y}_{t-i} + \boldsymbol{\beta} \boldsymbol{\alpha} \mathbf{y}_{t-1} + \boldsymbol{\Lambda} \mathbf{w}_t + \mathbf{u}_t$$

Estimation and testing for cointegrating rank remain exactly as before, although critical values of tests may be affected. Response of R20 to generalized one S.D. innovations

15.9 Common Trends and Cycles

Further implications of the presence of a linear trend in a CVAR are best analyzed by introducing the infinite-order vector polynomial $\mathbf{C}(B)$, defined such that

$$\begin{aligned} \mathbf{C}(B)\Pi(B) &= \nabla \mathbf{I}_n \\ \mathbf{y}_t &= \mathbf{y}_0 + \mathbf{b}_0 t + \mathbf{b}_1 \frac{t(t+1)}{2} + \mathbf{C}(B) \sum_{s=1}^t \mathbf{u}_s \\ &= \mathbf{y}_0 + \mathbf{b}_0 t + \mathbf{b}_1 \frac{t(t+1)}{2} + (\mathbf{C}(1) + \mathbf{C}^* \nabla) \sum_{s=1}^t \mathbf{u}_s \\ &= \mathbf{y}_0 + \mathbf{b}_0 t + \mathbf{b}_1 \frac{t(t+1)}{2} + \mathbf{C}(1) \mathbf{s}_t + \mathbf{C}^*(B) (\mathbf{u}_t - \mathbf{u}_0) \\ &= \mathbf{y}_0^* + \mathbf{b}_0 t + \mathbf{b}_1 \frac{t(t+1)}{2} + \mathbf{C}(1) \mathbf{s}_t + \mathbf{C}^*(B) \mathbf{u}_t \end{aligned}$$

where

$$\mathbf{y}_0^* = \mathbf{y}_0 - \mathbf{C}^*(B) \mathbf{u}_0, \quad \mathbf{s}_t = \sum_{s=1}^t \mathbf{u}_s$$

The inclusion of a linear trend in the VAR with $\mathbf{y}_t \sim I(1)$ thus implies a quadratic trend in the levels.

Consider then with the restriction $\mathbf{b}_1 = \mathbf{0}$ imposed and, for simplicity, initial values $\mathbf{y}_0 = \mathbf{u}_0 = \mathbf{0}$:

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

common trend-common cycle representation:

$$\mathbf{y}_t = \rho \tau_t + \mathbf{G} \tilde{\mathbf{c}}_t$$

Chapter 16 Compositional and Count Time Series

16.1 Constrained Time Series

When modeling such series, a "good" model should be unable to predict values which violate the known constraints, that is, the model should be "forecast coherent." Two examples of these types of series are considered in this chapter: (1) 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; and (2) "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

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

where $\mathbf{x}_i = (x_{1,i}, x_{2,i}, \dots, x_{T,i})'$, $i = 1, 2, \dots, D$, are such that $x_{t,i} > 0$ and $x_{t,1} + x_{t,2} + \cdots + x_{t,D} = 1$, $t = 1, 2, \dots, T$, that is, $\mathbf{x}_i > \mathbf{0}$ and

$$\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_D = \boldsymbol{\iota}$$

where $\boldsymbol{\iota} = [1 \quad 1 \quad \cdots \quad 1]'$ is a $T \times 1$ unit vector. The sub-matrix

$$\mathbf{X}^{(d)} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_d]$$

then lies in the d -dimensional simplex \mathcal{S}^d embedded in D -dimensional real space with

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

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

Several transformations have been proposed for doing this, the most popular being the additive log-ratio transformation which is defined as ²

$$\begin{aligned} \mathbf{Y} &= [\mathbf{y}_1 \quad \mathbf{y}_2 \quad \cdots \quad \mathbf{y}_d] = a_d(\mathbf{X}^{(d)}) \\ &= \left[\log\left(\frac{\mathbf{x}_1}{\mathbf{x}_D}\right) \quad \log\left(\frac{\mathbf{x}_2}{\mathbf{x}_D}\right) \quad \cdots \quad \log\left(\frac{\mathbf{x}_d}{\mathbf{x}_D}\right) \right] \end{aligned}$$

The inverse transformation, known as the additive-logistic, is

$$\begin{aligned} \mathbf{X}^{(d)} &= a_d^{-1}(\mathbf{Y}) = \left[\frac{\exp(\mathbf{y}_1)}{\mathbf{y}} \quad \frac{\exp(\mathbf{y}_2)}{\mathbf{y}} \quad \cdots \quad \frac{\exp(\mathbf{y}_d)}{\mathbf{y}} \right] \\ \mathbf{x}_D &= \frac{1}{\mathbf{y}} \end{aligned}$$

where

$$\mathbf{y} = 1 + \sum_{i=1}^d \exp(\mathbf{y}_i)$$

It is possible to avoid choosing a fill-up variable by using the centered log-ratio transformation, defined as

$$\mathbf{Z} = c_d \left(\mathbf{X}^{(d)} \right) = \begin{bmatrix} \log \left(\frac{\mathbf{x}_1}{g(\mathbf{X})} \right) & \log \left(\frac{\mathbf{x}_2}{g(\mathbf{X})} \right) & \cdots & \log \left(\frac{\mathbf{x}_D}{g(\mathbf{X})} \right) \end{bmatrix}$$

where

$$g(\mathbf{X}) = \begin{bmatrix} (x_{1,1} \times x_{1,2} \times \cdots \times x_{1,D})^{1/D} \\ (x_{2,1} \times x_{2,2} \times \cdots \times x_{2,D})^{1/D} \\ \vdots \\ (x_{T,1} \times x_{T,2} \times \cdots \times x_{T,D})^{1/D} \end{bmatrix}$$

is the vector of geometric means. Unfortunately, this has the disadvantage of introducing a non-singularity since $\mathbf{Z}\boldsymbol{\iota} = 0$.

16.3 Forecasting Compositional Time Series

A forecast for \mathbf{X}_{t+h} is then obtained as

$$\mathbf{X}_t(h) = a_d^{-1} (\mathbf{Y}_t(h))$$

From § 16.4 this forecast can be interpreted as an estimate of the geometric mean of $\mathbf{X}_{t+h} \sim L^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It is sometimes useful to provide forecasts for the ratios $x_{i,t+h}/x_{j,t+h}$ and these are given by

$$\left(\frac{x_i}{x_j} \right)_t(h) = \exp \left(y_{i,t}(h) - y_{j,t}(h) + \frac{1}{2} (\sigma_{ii,t}(h) - 2\sigma_{ij,t}(h) + \sigma_{jj,t}(h)) \right)$$

where $y_{i,t}(h)$ and $y_{j,t}(h)$ are the i th and j th elements of $\mathbf{Y}_t(h)$ and the $\sigma_{ij,t}(h)$ are the appropriate elements of $\boldsymbol{\Sigma}_t(h)$.

Under the normality assumption for \mathbf{Y}_t , a $100(1 - \alpha)\%$ confidence region for \mathbf{X}_{t+h} can be formed from:

$$\left(\mathbf{Y}_t(h) - \log \left(\frac{\mathbf{X}_{t+h}^{(d)}}{\mathbf{X}_{D,t+h}} \right) \right)' \boldsymbol{\Sigma}_t^{-1}(h) \left(\mathbf{Y}_t(h) - \log \left(\frac{\mathbf{X}_{t+h}^{(d)}}{\mathbf{X}_{D,t+h}} \right) \right) \leq \chi_\alpha^2(d)$$

where $\chi_\alpha^2(d)$ is the $100\alpha\%$ point of a $\chi^2(d)$ distribution, by mapping points from \mathcal{R}^d onto the simplex \mathcal{S}^d . Such a region, however, is probably only informative for $D \leq 3$, where graphical representations such as the ternary diagram are available.

16.4 Time Series Models for Counts: The IN-AR(1) Benchmark Model

The "benchmark" IN-AR(1) process is defined by the difference equation:

$$x_t = a \circ x_{t-1} + w_t$$

where the $x_t, t = 1, 2, \dots$, take on values in the set of nonnegative integers, $\mathcal{N} = \{0, 1, 2, \dots\}$. It is assumed that $0 \leq a < 1$ 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 . The process (16.6) is stationary and the discreteness of x_t is ensured by the **binomial thinning operation**

$$a \circ x_{t-1} = \sum_{i=1}^{x_{t-1}} y_{i,t-1}$$

where the $y_{i,t-1}$ are assumed to be i.i.d. Bernoulli random variables with

$$P(y_{i,t-1} = 1) = a$$

and

$$P(y_{i,t-1} = 0) = 1 - a$$

16.5 Other Integer-Valued ARMA Processes

The IN-MA(1) process is defined as

$$x_t = w_t + b \circ w_{t-1}$$

where $0 \leq b < 1$ and the thinning operation is defined analogously to (16.7) as

$$b \circ w_{t-1} = \sum_{i=1}^{w_{t-1}} y_{i,t-1}$$

The autocorrelation function (ACF) of x_t is now

$$\rho_k = \begin{cases} \frac{b\sigma_w^2}{b(1-b)\mu_w + (1+b^2)\sigma_w^2} & \text{for } k = 1 \\ 0 & \text{for } k > 1 \end{cases}$$

A natural extension of (16.6) is to the IN-AR(2) process

$$x_t = a_1 \circ x_{t-1} + a_2 \circ x_{t-2} + w_t$$

16.6 Estimation of Integer-Valued ARMA Models

the "bias-corrected Yule-Walker" estimate (recall § 3.20):

$$\hat{a} = \frac{1}{T-3}(Tr_1 + 1)$$

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

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

By extension, the Yule-Walker estimates of a_1 and a_2 in the Po-IN-AR(2) model are given by (again see § 3.20):

$$\hat{a}_1 = \frac{r_1(1-r_2)}{1-r_1^2} \quad \hat{a}_2 = \frac{r_2-r_1^2}{1-r_1^2}$$

and it then follows that an estimate of λ is given by

$$\hat{\lambda}_2 = (1 - \hat{a}_1 - \hat{a}_2)\bar{x}$$

16.7 Testing for Serial Dependence in Count Time Series

After an extensive Monte Carlo simulation exercise, Jung and Tremayne (2003) suggest focusing attention on three tests. The first is the score test, $S^* = \sqrt{T}r_1 \sim N(0, 1)$, 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. The other two tests are

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

and

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

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

16.8 Forecasting Counts

The approach to forecasting taken in Chapter 7, An Introduction to Forecasting with Univariate Models, is based on the conditional expectation, that is,

$$f_{T,h} = E(x_{T+h} | x_T, x_{T-1}, \dots, x_1)$$

and is known to yield MMSE forecasts (cf. §7.2). The conditional expectation of the IN-AR(1) model is given in § 16.10, so that

$$\begin{aligned} f_{T,1} &= ax_T + \mu_w \\ f_{T,2} &= af_{T,1} + \mu_w = a^2 x_T + (1+a)\mu_w \end{aligned}$$

and

$$f_{T,h} = a^h x_T + (1 + a + a^2 + \dots + a^{h-1})\mu_w$$

This has the following (fairly) tractable expression for the IN-AR(1) model under a Poisson innovation assumption:

$$p_h(x | x_T) = x_T! \exp \left(-\lambda \left(\frac{(1-a^h)}{(1-a)} \right) \right) C_h(x_T, x) \quad x = 0, 1, 2, \dots$$

where

$$C_h(x_T, x) = \sum_{k=0}^m \frac{a^{kh} (1-a^h)^{x_T-k} \lambda^{x-k}}{k!(x-k)! (|x_T - k|)!} \quad m = \min(x_T, x)$$

16.9 Intermittent and Nonnegative Time Series

When a count series contains many zeros, it is sometimes referred to as being intermittent, as when a stock inventory is only occasionally demanded.

Chapter 17 State Space Models

17.1 Formulating State Space Models

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**: see, e.g., Harvey, 1989, Chapters 3 and 4; Hamilton, 1994, Chapter 13; or Durbin and Koopman, 2012, for full textbook treatments). Although there are various specifications of the SSF, a popular version has the measurement equation taking the form:

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

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 $V(\varepsilon_t) = h_t$. In general, the elements of the $m \times 1$ **state vector** $\boldsymbol{\alpha}_t$ are unobservable, but are assumed to be generated by the transition equation

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

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

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

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

and

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

If the system matrices do not change over time, the model is said to be **time-invariant or time-homogeneous**, a special case of which are stationary models.

1. The AR(2) process $x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \eta_t$ has, on defining the state vector as

$$\boldsymbol{\alpha}_t = \begin{bmatrix} x_t \\ \phi_2 x_{t-1} \end{bmatrix}$$

the SSF

$$\begin{aligned} x_t &= [1 \ 0] \boldsymbol{\alpha}_t + \phi_0 \\ \boldsymbol{\alpha}_t &= \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix} \boldsymbol{\alpha}_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_t \end{aligned}$$

i.e., the system matrices are, with $m = 2$ and $g = 1$, so that $\boldsymbol{\eta}_t = \eta_t$,

Alternatively, we could define

$$\begin{aligned} \mathbf{z}'_t &= \mathbf{R}'_t = [1 \ 0] \quad d_t = \phi_0 \quad h_t = 0 \quad \mathbf{c}'_t = [0 \ 0] \\ \mathbf{T}_t &= \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix} \end{aligned}$$

$$\boldsymbol{\alpha}_t^* = \begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}$$

and the SSF

$$x_t = [1 \ 0] \boldsymbol{\alpha}_t^* + \phi_0$$

$$\boldsymbol{\alpha}_t^* = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \boldsymbol{\alpha}_{t-1}^* + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_t$$

2. The MA(1) process $x_t = \eta_t + \theta\eta_{t-1}$ has SSF

$$x_t = [1 \ 0] \boldsymbol{\alpha}_t$$

$$\boldsymbol{\alpha}_t = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \boldsymbol{\alpha}_{t-1} + \begin{bmatrix} 1 \\ \theta \end{bmatrix} \eta_t$$

on defining $\boldsymbol{\alpha}_t = [x_t \ \theta\eta_{t-1}]'$.

3. Consequently, the general ARMA(p, q) process, written in the form:

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_m x_{t-m} + \eta_t + \theta_1 \eta_{t-1} + \cdots + \theta_{m-1} \eta_{t-m+1}$$

where $m = \max(p, q + 1)$, has the SSF

Other models discussed in earlier chapters may also be cast in SSF form. The Muth/simple exponential smoothing model (cf. § 8.1 and § 9.15), when written in the notation of this chapter, viz.,

$$x_t = \alpha_t + \varepsilon_t \quad E(\varepsilon_t) = 0 \quad V(\varepsilon_t) = h$$

$$\alpha_t = \alpha_{t-1} + \eta_t \quad E(\eta_t) = 0 \quad V(\eta_t) = q = \kappa h$$

is seen to be an SSF with $m = g = 1$, $d_t = \mathbf{c}_t = 0$ and $z_t = \mathbf{T}_t = \mathbf{R}_t = 1$.

Similarly, the Holt-Winters local linear trend model can be written as:

$$x_t = \alpha_{1,t} + \varepsilon_t$$

$$\alpha_{1,t} = \alpha_{1,t-1} + \alpha_{2,t-1} + \eta_{2,t}$$

$$x_t = [1 \ \mathbf{0}'_{m-1}] \boldsymbol{\alpha}_t$$

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \phi_1 & \vdots \\ \phi_2 & \vdots \\ \vdots & \vdots \\ \phi_{m-1} & \vdots \\ \dots & \dots \\ \dots & \mathbf{0}'_{m-1} \end{bmatrix} \boldsymbol{\alpha}_{t-1} + \begin{bmatrix} 1 \\ \phi_m \\ \vdots \\ \vdots \\ \theta_{m-1} \end{bmatrix} \eta_t$$

$$\alpha_{2,t} = \alpha_{2,t-1} + \eta_{3,t}$$

and is in SSF form with $m = g = 2$, $z_t = [1 \ 0]$, $\mathbf{R}_t = \mathbf{I}_2$ and

$$\mathbf{T}_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

17.2 The Kalman Filter

Once a model has been put into state space form, several important algorithms may be applied. Central to these is the **Kalman (-Bucy)** filter.

17.3 ML Estimation and the Prediction

Consider, then, the state space model of (17.1) and (17.2). Let α_{t-1} be the optimal estimator of α_{t-1} based on observations up to and including x_{t-1} , that is, $\alpha_{t-1} = E_{t-1}(\alpha_{t-1} | \mathbf{x}_{t-1})$, where $\mathbf{x}_{t-1} = \{x_{t-1}, x_{t-2}, \dots, x_1\}$, and let

$$\mathbf{P}_{t-1} = E(\alpha_{t-1} - \alpha_{t-1})(\alpha_{t-1} - \alpha_{t-1})'$$

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

$$a_{t|t-1} = \mathbf{T}_t a_{t-1} + c_t$$

and

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

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

$$\alpha_t = a_{t|t-1} + \mathbf{P}_{t|t-1} z_t f_t^{-1} (x_t - z'_t a_{t|t-1} - d_t)$$

and

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

where

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

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

$$\alpha_{t+1|t} = (\mathbf{T}_{t+1} - \mathbf{K}_t z'_t) \alpha_{t|t-1} + \mathbf{K}_t x_t + \mathbf{c}_{t+1} - \mathbf{K}_t d_t$$

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

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

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

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

17.4 Error Decomposition

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

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

Clearly, the principal characteristic of a time series model is that the observations are not independent so that is not directly applicable. Instead, the definition of a conditional probability density function must be used to write the joint density as:

$$\mathcal{L}(\mathbf{x} : \boldsymbol{\psi}) = \prod_{t=1}^T p(x_t | \mathbf{x}_{t-1})$$

where $p(x_t | \mathbf{x}_{t-1})$ denotes the distribution of x_t conditional on the information set available at time $t - 1$.

If the errors and initial state vector in (17.1) are normally distributed, then the distribution of x_t , conditional on \mathbf{x}_{t-1} , will itself be normal, with the mean and variance of this conditional distribution being given by the Kalman filter. Conditional on \mathbf{x}_{t-1} , α_t is normally distributed with mean $\mathbf{a}_{t|t-1}$ and covariance matrix $\mathbf{P}_{t|t-1}$. If the measurement equation is written as:

$$x_t = z'_t \mathbf{a}_{t|t-1} + z'_t (\alpha_t - \mathbf{a}_{t|t-1}) + d_t + \varepsilon_t$$

then the conditional distribution of x_t is normal with mean

$$E_{t-1}(x_t) = \hat{x}_{t|t-1} = z'_t \mathbf{a}_{t|t-1} + d_t$$

and variance f_t . The likelihood function can then be written immediately 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 is also known as the ***prediction error decomposition*** form of the likelihood function.

17.5 Prediction and Smoothing

Applying the Kalman filter to (17.1) and (17.2) yields \mathbf{a}_T , the MMSE of α_T , based on all T observations. In addition, it gives:

$$\mathbf{a}_{T+1|T} = \mathbf{T}_{T+1} \mathbf{a}_T + \mathbf{c}_{T+1}$$

together with the one-step ahead prediction

$$\hat{x}_{T+1|T} = z'_{T+1} \mathbf{a}_{T+1|T} + d_{T+1}$$

We may also consider the problem of multistep prediction, that of making predictions of future observations at times $T + 2, T + 3, \dots$. Extending (17.13) by substituting repeatedly into the transition equation at time $T + l$ and taking conditional expectations at time T yields

$$\mathbf{a}_{T+l|T} = \left[\prod_{j=1}^l \mathbf{T}_{T+j} \right] \mathbf{a}_T + \sum_{j=1}^{l-1} \left[\prod_{i=j+1}^l \mathbf{T}_{T+i} \right] \mathbf{c}_{T+j} + \mathbf{c}_{T+l}$$

The MMSE of x_{T+l} is then obtained directly from $\mathbf{a}_{T+l|T}$ as

$$\hat{x}_{T+l} = z'_{T+l} \mathbf{a}_{T+l|T} + d_{T+l}$$

with the accompanying MSE

$$\text{MSE}(\hat{x}_{T+l|T}) = z'_{T+l} \mathbf{P}_{T+l|T} z_{T+l} + h_{T+l}$$

The easiest way to calculate $\mathbf{a}_{T+l|T}$ and $\mathbf{P}_{T+l|T}$, which are required to evaluate (17.14) and (17.15), is to repeatedly apply the Kalman filter prediction equations (17.4) and (17.5) to obtain

$$\mathbf{a}_{T+l|T} = \mathbf{T}_{T+l} \mathbf{a}_{T+l-1|T} + \mathbf{c}_{T+l}$$

and

$$\mathbf{P}_{T+l|T} = \mathbf{T}_{T+l}\mathbf{P}_{T+l-1|T}\mathbf{T}'_{T+l} + \mathbf{R}_{T+l}\mathbf{Q}_{T+l}\mathbf{R}'_{T+l}$$

While there are several forms of smoothing, attention is concentrated here on ***fixed-interval smoothing***. This is an algorithm that consists of a set of recursions which start with the final quantities, \mathbf{a}_T and \mathbf{P}_T , given by the Kalman filter and work backward. These equations are:

$$\mathbf{a}_{t|T} = \mathbf{a}_T + \mathbf{P}_t^* (\mathbf{a}_{t+1|T} - \mathbf{T}_{t+1}\mathbf{a}_t)$$

and

$$\mathbf{P}_{t|T} = \mathbf{P}_t + \mathbf{P}_t^* (\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|T})\mathbf{P}_t^{*\prime}$$

where

$$\mathbf{P}_t^* = \mathbf{P}_t \mathbf{T}'_{t+1} \mathbf{P}_{t+1|t}^{-1} \quad t = T-1, \dots, 1$$

with $\mathbf{a}_{T|T} = \mathbf{a}_T$ and $\mathbf{P}_{T|T} = \mathbf{P}_T$.

17.6 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 \mathbf{X}_t of observed series by generalizing the measurement equation to

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

where \mathbf{Z}_t is an $N \times m$ matrix, \mathbf{d}_t is an $N \times 1$ vector, and $\boldsymbol{\varepsilon}_t$ is an $N \times 1$ vector with $E(\boldsymbol{\varepsilon}_t) = \mathbf{0}$ and $V(\boldsymbol{\varepsilon}_t) = \mathbf{H}_t$, an $N \times N$ covariance matrix.

Chapter 18 Some Concluding Remarks

1. a set of robust and useful techniques that we can call upon to analyze time series that occur in our own research fields
2. to begin, wherever possible, with a general model and then test plausible restrictions with the aim of moving toward a simpler model that has empirical support
3. an appreciation of the principle that the properties a time series displays will impact upon the behavior to be expected from it.
4. As well as incorporating and testing relevant theory considerations, "institutional" knowledge can also be important when analyzing data and time series are no exceptions to this.