Univariate Time Series Analysis

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Univariate Time Series Analysis

In this lecture, we will present a brief introduction to univariate time series analysis. We will discuss moving average (MA), autoregressive (AR), and autoregressive moving average (ARMA) time series models, examine their equivalence relationships, and discuss forecasting and estimation in the context of these models.

We will also discuss the Wold theorem and its implications for statistical reduced-form representations of the Roll model in the form of MA and AR models.

Overview

- The Roll model described in the previous lecture is a simple structural model with a clear mapping to parameters that are easily estimated using the variance and autocovariance of price changes.

- There are many interesting questions, however, that go beyond basic parameter estimation. We might want to *forecast* prices beyond the end of our data sample, or identify the (unobserved) efficient price series m_t underlying our data.

- Furthermore, when we suspect that the structural model is mis-specified, we might prefer to make assumptions about the data, rather than about the model.

Overview

- To address these issues, in this lecture we'll examine the Roll model from a different point of view.
- Whereas in the previous lecture we took a structural economic perspective, here we adopt a more data-oriented statistical reduced-form approach.
- In the process of going back and forth between the structural and statistical representations, we will illustrate econometric techniques that are useful in more general situations.
- We will begin by describing some useful general properties of the time series and the proceed to moving average and autoregressive models, finally describing forecasting and estimation.

Stationarity and Ergodicity

- Much statistical inference relies on the **law of large numbers** (LLN) and the **central limit theorem** (CLT).
- These results establish the limiting properties of estimators as the sample size increases.
- The usual forms of these theorems apply to data samples consisting of independent observations, but time series data are by nature *dependent*.
- To maintain the strength of the LLN and CLT when independence does not hold, we rely on alternative versions of these results that assume stationarity and ergodicity.

Definition: Covariance Stationarity

A time series $\{x_t\}$ with constant mean $E[x_t] = \mu$, and autocovariances $Cov(x_t, x_{t-k}) = \gamma_k$ that do not depend on t is said to be *covariance stationary*.

Stationarity and Ergodicity

Definition: Strict Stationarity

A time series $\{x_t\}$ for which all joint density functions of the form $f(x_t), f(x_t, x_{t+1}), \ldots, f(x_t, x_{t+1}, x_{t+2}), \ldots$ do *not* depend on t is strictly stationary.

Example: the price **changes** Δp_t implied by the Roll model, are covariance stationary: $E[\Delta p_t] = 0 \; and \; Cov(\Delta p_t, \Delta p_{t-k}) = \gamma_k$.

- The price **levels** are **not** covariance stationary; among other things, variance $Var(p_t)$ increases with t).
- Covariance stationarity of Δp_twould also be a violated if we replaced the homoscedasticity assumption $E[u_t^2] = \sigma_u^2$ with a time-dependent (heteroscedasticity) feature.

Memory

- We sometimes describe a sequence of independent observations by saying that an observation carries *no memory* of observations earlier in the sequence.
- This is too restrictive for time series analysis.
- We typically assume instead that the effects of earlier observations *decay* and *die out* with the passage of time.

Stationarity and Ergodicity

Definition: Ergodicity

A time series is *ergodic* if its local stochastic behavior is *independent* of initial conditions.

An ergodic process eventually "forgets" where it started. The *price level* in the Roll model is *not* ergodic: the randomness in the level is cumulative over time. But the *price changes* are ergodic: Δp_t is independent of Δp_{t-k} for $k \geq 2$.

Example:

Nonergodicity could be introduced by positing $m_t = m_{t-1} + u_t + \mathrm{z}$, where z is a zero-mean (random) variable drawn *once* at time zero.

- Ergodicity, in the sense of dependence on the initial conditions, may be an important attribute of market mechanisms.
- In the long run, we would expect security prices to reflect *fundamentals.*
- A trade mechanism that induces persistent price components might impair its adjustments.

Moving Average (MA) Models

We will often assume that a time series like $\{\Delta p_t\}$ is covariance stationary, and now we turn to various ways in which the series can be represented. We start with a white noise process:

Definition: White Noise Process

A time series $\{\varepsilon_t\}$ where $E[\varepsilon_t] = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2$ and $Cov(\varepsilon_t, \varepsilon_{t-k}) = 0$ for $k \neq t$ is called a **white noise** process.

- A white noise process is obviously not covariance stationary.
- In many economic settings, is convenient and plausible to assume that $\{\varepsilon_t\}$ are strictly stationary and even normally distributed, but we will avoid those assumptions.
- White noise processes are convenient building blocks for constructing dependent time series.
- One such construction is the moving average (MA) model.

The moving average model of order one MA(1) is:

$$x_t = \varepsilon_t + \theta\varepsilon_{t-1} \qquad \text{(EQ 1)}$$

Moving Average (MA) Models

The white noise driving a time series model is variously termed the *disturbance*, *error* or *innovation* series. From the statistical viewpoint, they all amount to the same thing. The economic interpretations and connotations, however, vary.

- When randomness is being added to a non-stochastic dynamic structural model, the term “disturbance” suggests a *shock* to which the system subsequently *adjusts*.

- When *forecasting* is the main concern, “error” conveys a sense of *discrepancy* between the observed value and the model prediction.

- “Innovation” is the word that is most loaded with economic connotations. The innovation is what the econometrician learns about the price at time *t* (beyond what is known from prior observations). Moving forward in time, it is the update to econometrician’s *information set*.

Moving Average (MA) Models

The Δp_t in the Roll model have the property that the autocovariances are zero beyond lag one. The MA(1) model in Equation 1 also has this property. For this process, the variance and first order autocovariances are:

$$\gamma_0 = (1+\theta^2)\sigma_\varepsilon^2,$$
$$\gamma_1 = \theta\sigma_\varepsilon^2, \text{ and}$$
$$\gamma_k = 0 \text{ for } k > 1 \qquad \text{(EQ 2)}$$

More generally, the moving average model of order K is:

$$x_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_K\varepsilon_{t-K} \qquad \text{(EQ 3)}$$

The MA(K) process is *covariance stationary* and has the property that $\gamma_j = 0$ for $j > K$. If we let $K = \infty$, we arrive at the infinite order moving average process MA(∞):

$$x_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots \qquad \text{(EQ 4)}$$

Moving Average (MA) Models

Representations of the Roll Model

- If we believe that the $\{\Delta p_t\}$ are generated by the Roll model (a *structural economic* model), can we assert that a corresponding moving average model (a *statistical* model) exists?
- By playing around with the θ and σ_ε^2 parameters in the MA(1) model, we can obviously *match* the variance and the first-order autocovariance of the structural Δp_t process.
- But this is not quite the same thing as claiming that the full joint distribution of the Δp_t realizations generated by the structural model could also be generated by an MA(1) model.
- Moreover, there's a good reason for suspecting it shouldn't be possible. The structural model has two (uncorrelated) sources of randomness; the MA(1) has only one source of randomness, ε_t.

Moving Average (MA) Models

Is the existence of a MA(1) representation an important issue? Why can't we simply limit the analysis to the structural model and avoid the questions of alternative representations? There are several answers.

- In the first place, the full structural model involves unobservable variables. The econometrician observes neither u_t nor q_t, so he or she doesn't know the efficient price.
- The moving average representation is a useful tool for constructing an estimate of the efficient price as well as for forecasting.
- Moreover, a moving average representation may be valid even if the structural model is mis-specified.

Moving Average (MA) Models

Fortunately, an MA(1) representation for the price changes in the Roll model *does* exist. In this assertion we rely on the Wold theorem. The Wold theorem states that any zero-mean covariance stationary process $\{x_t\}$ can be represented in the form:

$$x_t = \sum_{j=0}^{\infty} \theta_j \, \varepsilon_{t-j} + \kappa_t$$

(EQ 5)

where $\{\varepsilon_t\}$ is a zero-mean white noise process, $\theta_0 = 1$(normalization) and $\sum_{j=0}^{\infty} \theta_j^2 < \infty$.

κ_t is a linearly deterministic process, which in this context means that it can be predicted arbitrarily well by a linear projection (possibly of infinite order) on past observations of x_t.

For a purely stochastic series $\kappa_t = 0$, we are left with a moving average representation.

Moving Average (MA) Models

Theorem (Ansley, Spivey, and Wrobleski (1977))

If a covariance stationary process has *zero autocorrelations* of all orders *higher* than *K*, then it possesses a moving average representation of order *K*.

This theorem allows us to assert that an MA(1) representation exists for the Roll model.

- Empirical market microstructure analyses often stretch the Wold theorem. The structural models are often *stylized* and *under-identified* (i.e. we can't estimate all the parameters).
- The data are frequently *non-Gaussian* (like the trade indicator variable in the Roll model).
- Covariance stationarity of the observations (possibly after a transformation) is often *tenable* working assumption. For many purposes, as we'll see, it is enough.

Moving Average (MA) Models

- In the previous lecture we derived autocorrelations of the Roll model in terms of the structural parameters (c and σ_u^2).
- The parameters of the corresponding MA(1) model in Equation 1 above are θ and σ_ε^2.
- The MA(1) has variance and autocovariances $\gamma_0 = (1+\theta^2)\sigma_\varepsilon^2$, and $\gamma_1 = \theta\sigma_\varepsilon^2$.
- From the autocovariances (or estimates thereof), we may compute the moving average parameters:

$$\theta = \frac{\gamma_0 - \sqrt{\gamma_0^2 - 4\gamma_1^2}}{2\gamma_1} \text{ and } \sigma_\varepsilon^2 = \frac{\gamma_0 - \sqrt{\gamma_0^2 - 4\gamma_1^2}}{2}$$

(EQ 6)

- This is actually one of the *two* solutions, the so-called *invertible solution*. It has the property that $|\theta| < 1$, the relevance of which will shortly become clear.
- The other (non-invertible) solution is $\{\theta^*, \sigma_\varepsilon^{2*}\}$, where $\theta^* = 1/\theta$ and $\sigma_\varepsilon^{2*} = \theta^2\sigma_\varepsilon^2$. For the noninvertible solution, $|\theta^*| > 1$.

Autoregressive (AR) Models

- A moving average model expresses the **current** realization in terms of ***current*** and ***lagged disturbances***. These are not generally observable.
- For many purposes (particularly forecasting) it is useful to express the **current** realizations in terms of the ***past realizations***.
- This leads to the autoregressive form of the model.

Autoregressive (AR) Models

To develop this for the MA(1) case, note that we can rearrange:

$\Delta p_t = \varepsilon_t + \theta\varepsilon_{t-1}$ as:

$\varepsilon_t = \Delta p_t - \theta\varepsilon_{t-1}$.

This gives us a backward recursion for ε_t :

$\varepsilon_{t-1} = \Delta p_{t-1} - \theta\varepsilon_{t-2}$, $\varepsilon_{t-2} = \Delta p_{t-2} - \theta\varepsilon_{t-3}$, and so on. Using this backward recursion gives:

$$\Delta p_t = \theta\left(\Delta p_{t-1} - \theta\big(\Delta p_{t-2} - \theta(\Delta p_{t-3} - \cdots)\big)\right) + \varepsilon_t$$

(EQ 7)

Leading to:

$$\Delta p_t = \theta\Delta p_{t-1} - \theta^2\Delta p_{t-2} + \theta^3\Delta p_{t-3} + \cdots + \varepsilon_t$$

(EQ 8)

This is the autoregressive form: Δp_t is expressed as a linear function of its own lagged values and the current disturbance. Although the moving average representation is of order one, the autoregressive representation is of infinite order.

If $|\theta| < 1$, then the autoregressive representation is convergent: the coefficients of the lagged Δp_t converge to zero. Intuitively, the effects of the lagged realizations eventually die out.

Autoregressive (AR) Models

Definition (Invertible MA Representation): When a convergent autoregressive representation *exists*, the moving average representation is said to be *invertible*.

Convergence is determined by the *magnitude* of θ. The condition $|\theta| < 1$ thus defines the invertible solution for the MA(1) parameters.

To move between the moving average and autoregressive representations, it is often convenient to use the *lag operator* L. It is defined by the relation $Lx_t = x_{t-1}$. Multiple applications work in a straightforward fashion, $L^2 x_t = x_{t-2}$, $L^3 x_t = x_{t-3}$ etc. The operator can also generate "leads", e.g. $L^{-1} x_t = x_{t+1}$, $L^{-2} x_t = x_{t+2}$ etc. Using the lag operator, the *moving average representation* for Δp_t is:

$$\Delta p_t = \left(\theta L - \theta^2 L^2 + \theta^3 L^3 - \ldots\right)\Delta p_t + \varepsilon_t \qquad \text{(EQ 9)}$$

Autoregressive (AR) Models

We derived this by recursive substitution. But there is an *alternative* construction that is particularly useful when the model is complicated. Starting from the moving average representation, we may write

$$\Delta p_t = (1+\theta L)\varepsilon_t \Rightarrow (1+\theta L)^{-1}\Delta p_t = \varepsilon_t \qquad \text{(EQ 10)}$$

where we have essentially treated the lag operator term as an *algebraic* quantity. If L were a variable and $|\theta| < 1$, we could construct a *series expansion* of the left-hand side around $L = 0$. This expansion through the third order would be

$$\left(1-\theta L+\theta^2 L^2-\theta^3 L^3+O(L^4)\right)\Delta p_t = \varepsilon_t \qquad \text{(EQ 11)}$$

where $O(L^4)$ represents the higher order terms. This can be rearranged to obtain the autoregressive representation in Equation 9.

Autoregressive (AR) Models

In summary, we have modeled a time series by assuming covariance stationarity, proceeding to a moving average representation (via Wold theorem), and finally to the autoregressive representation. The last two representations are *equivalent*, but in any particular problem, one might be considerably *simpler* than the other. For example, the Roll model is a moving average of order one, but the autoregressive representation is of infinite order.

Sometimes, though, the autoregressive representation is the simpler one. An autoregressive representation of order one AR(1) has the form:

$$x_t = \varphi x_{t-1} + \varepsilon_t \qquad \text{(EQ 12)}$$

or in terms of the lag operator L:

$$(1 - \varphi L)x_t = \varepsilon_t \qquad \text{(EQ 13)}$$

Autoregressive (AR) Models

The equivalent *moving average* form of this is:

$$x_t = (1-\varphi L)^{-1}\varepsilon_t = (1+\varphi L+\varphi^2 L^2+\ldots)\varepsilon_t$$

$$= \varepsilon_t + \varphi\varepsilon_t + \varphi^2 \varepsilon_t{}^2 + \ldots \qquad \text{(EQ 14)}$$

Here, we have used a power *series expansion* of $(1-\varphi L)^{-1}$. Recursive substitution would give the same result. The *moving average* representation is of *infinite order*.

Forecasting

A crucial calculation in the agents' trading decision is their forecast of the security's future value. It is convenient to construct these forecasts by taking *expectations* of MA and AR representations, but there is an important qualification. The assumption of covariance stationarity suffices only to characterize a *restricted* form of the expectation. An expectation, e.g. $E[x_t \mid x_{t-1}, x_{t-2}, \ldots]$, generally involves the *full joint distribution* $f(x_t, x_{t-1}, x_{t-2}, \ldots)$, not just the means and covariances. Considerable simplification results, however, if we approximate the true expectation by *linear* functions of the conditioning arguments, that is

$$E[x_t \mid x_{t-1}, x_{t-2}, \ldots] \approx \alpha_0 + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots \qquad \text{(EQ 15)}$$

This approximate expectation is technically a *linear projection*. When the difference is important it will be denoted by E^* to distinguish it from the true expectation. The following material summarizes the results on *linear forecasting*.

Forecasting

The technique of linear projection is especially compatible with AR and MA representations because AR and MA representations have no more or no less information than is needed to compute the projection. It is quite conceivable that a more complicated forecasting scheme, for example, one involving nonlinear transformations of $\{x_{t-1}, x_{t-2}, \ldots\}$ might be better (have smaller forecasting error) than the linear projection. But such a forecast could not be computed directly from the AR or MA representation. More structure would be needed.

We'll first consider the *price forecast* in the Roll model. Suppose that we *know* θ and have full (infinite) price history up to the time *t*, $\{p_t, p_{t-1}, p_{t-2}, \ldots\}$. Using the *autoregressive representation*, we can *recover* the innovations series $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$. Then:

$$E^*\left[\Delta p_{t+1} \mid p_t, p_{t-1}, p_{t-2}, \ldots\right] = E^*\left[\varepsilon_{t+1} + \theta\varepsilon_t \mid p_t, p_{t-1}, p_{t-2}, \ldots\right] = \theta\varepsilon_t \qquad \text{(EQ 16)}$$

Forecasting

Therefore, the *forecast* of the *next* period's price is:

$$f_t = E^*\left[p_{t+1} \mid p_t, p_{t-1}, p_{t-2}, \ldots\right] = p_t + \theta\,\varepsilon_t \qquad \text{(EQ 17)}$$

We can ask how does f_t *evolve*:

$$\Delta f_t = f_t - f_{t-1} = p_t + \theta\,\varepsilon_t - \left(p_{t-1} + \theta\,\varepsilon_{t-1}\right) = \left(p_t - p_{t-1}\right) + \theta\left(\varepsilon_t - \varepsilon_{t-1}\right)$$

$$= \left(\varepsilon_t + \theta\,\varepsilon_{t-1}\right) + \theta\left(\varepsilon_t - \varepsilon_{t-1}\right) = (1+\theta)\varepsilon_t \qquad \text{(EQ 18)}$$

That is, the *forecast revision* is a constant *multiple* of the innovation. The innovation process is uncorrelated, so the forecast revision is also *uncorrelated*.

Forecasting

Now we raise a more difficult question. A Martingale has uncorrelated increments, so f_t *might be* a martingale. Can we assert that $f_t = m_t$, that is, have we identified the true implicit efficient price? It turns out that there is a bit of problem. If $f_t = m_t$, then $p_t = f_t + cq_t$ and $\Delta p_t = \Delta f_t + c\Delta q_t$. But this implies

$$\Delta p_t = \varepsilon_t + \theta\varepsilon_{t-1} = (1+\theta)\varepsilon_t + c\Delta q_t \Leftrightarrow -\theta(\varepsilon_t - \varepsilon_{t-1}) = c\Delta q_t \qquad \text{(EQ 19)}$$

In other words, all the randomness in the model is attributable to the q_t. But this is *structurally incorrect*: We know that changes in the efficient price, u_t, also contribute to the ε_t. Thus, we have *not* identified the efficient price m_t. It will later be shown that

$$f_t = E^*[m_t \mid p_{t,} p_{t-1}, p_{t-2}, \ldots] \qquad \text{(EQ 20)}$$

that is, that f_t is the *linear projection* of m_t on the *past prices*.

Estimation

- In practice, the Roll model parameters are usually estimated as transformations of the estimated variance and first order autocovariance of the price changes.
- It is not uncommon, however, for the estimated first-order autocovariance to be *positive*. This can be due to estimation error, even though the model is correctly specified.
- More generally, MA and AR representations can be estimated using a wide variety of approaches.
- The MA parameters can be obtained from the autocovariances (by solving the set of equations and requiring that the solution be invertible).
- MA model can be estimated via *maximum likelihood* (assuming a particular distribution for disturbances).
- The MA representation can be obtained by *numerically inverting* the AR representation.

Estimation

The autoregressive representation can often be conveniently estimated using ordinary least squares (OLS).

- The basic representation for consistency of the OLS estimate is that the residuals are uncorrelated with the regressors.
- This is true in Equation 8 because the $\{\varepsilon_t\}$ are serially uncorrelated and the regressors (lagged price changes) are linear functions of prior realizations of ε_t.
- For example, $\Delta p_t = \varepsilon_{t-1} + \theta\varepsilon_{t-2}$ is uncorrelated with ε_t.

Estimation

Microstructure data often present particular challenges.

- Samples often contain embedded breaks. In a sample of intra-day trade prices that spans multiple days, for example, the closing price on one day and the opening price on the following day will appear successively.
- The overnight price change between these observations, though, will almost certainly have different properties than the intraday price changes.
- If the goal is modeling the latter, the overnight price changes should be dropped. This is often accomplished by inserting missing values in the series at day breaks.

Estimation

- A related issue concerns *lagged values* realized *before* the *start* of the sample. In an autoregressive representation if *t* is the *first* observation of the sample, none of the *lagged* values on the right-hand side are known.
- Most *non-microstructure* applications take the perspective that the start of sample simply represents the beginning of the record for a process that was *already unfolding*.
- The correct estimation approach is then unconditional, that is, the lagged missing values are viewed as *unknown* but *distributed* in accordance with the *model*.
- In many *microstructure* situations, though, the data *begin* right at the start of trading process.
- There is *no* prior unobserved evolution of the trading process. In these cases, conditional estimation, wherein the *missing* lagged disturbances are set to *zero*, is more defensible.

Limitations of Linear Time Series Models

- This lecture has reviewed the elements of linear time-series analysis.
- The development begins with covariance stationarity, which is a plausible and minimal working assumption in many modeling situations.
- Using the Wold theorem, this leads to a moving average model, then to a vector autoregression, and finally to a forecasting procedure.
- These are powerful results, but to maintain a balanced perspective, it is now necessary to consider some of the framework's limitations.

Limitations of Linear Time Series Models

- The characterization of a time series offered by the linear models is *not complete*. The models *do not* fully describe the *data-generating process*- they *do not* specify how we should computationally simulate the process.
- The disturbances in MA and AR models are *serially uncorrelated*, but may be *serially dependent*. This bears directly on the structural interpretations of these models.
- The MA and AR representations of a discretely valued process such as q_t in the Roll model are essentially *linear models* of *limited dependent variables*.

Limitations of Linear Time Series Models

Linear time-series analysis nevertheless retains strength and utility.

- It provides *logically coherent* and *computationally simple* tools for describing first-order *dynamics*, *forecasting*, and forming *expectations*.
- The underlying assumptions are *minimal* (chiefly covariance stationarity), so the analyses may be more *robust* to misspecification than more refined models.
- The representations are compatible with a wide range of structural models and so are relatively easy to illustrate and interpret.
- In short, they are useful aids in developing intuitions of how financial markets work.

Appendix: The Wold Theorem

- ARMA models comes with *restrictions* which ensure they are models for *covariance stationary* time series.
- In an ARMA model the data are a linear combination of current and past one-step ahead forecast errors, with weights that *decay* at a *geometric rate*.
- Here we consider the class of covariance stationary processes and ask whether ARMA models are a strict *subset* of that class.
- We start from the assumption that a process is covariance stationary and we study the projections of the process onto its *current* and *past one-step-ahead forecast errors (*the 'purely indeterministic/stochastic part' of the process) and a projection error (the 'purely deterministic part').
- This is know as the Wold Representation Theorem.

Appendix: The Wold Theorem

The Wold Representation Theorem: Suppose that $\{x_t\}$ is a *covariance stationary* process with $E\,x_t = 0$ and covariance function $\gamma_k = E\,x_t x_{t-k}$ for all k. Then

$$x_t = \sum_{j=0}^{\infty} d_j \varepsilon_{t-j} + \eta_t \qquad \text{(A1)}$$

Where

$$d_0 = 1, \quad \sum_{j=0}^{\infty} d_j^2 < \infty, \quad E\,\varepsilon_t^2 = \sigma_\varepsilon^2, \quad E\,\varepsilon_t \varepsilon_s = 0 \text{ for } t \neq s,$$

$$E\,\varepsilon_t = 0, \quad E\,\eta_t \varepsilon_s = 0 \text{ for } t \neq s, \quad P\{\eta_{t+s} \mid x_{t-1}, x_{t-2}, \ldots\} = \eta_{t+s} \quad s \geq 0 \qquad \text{(A2)}$$

Appendix: The Wold Theorem

Here P denotes a *linear projection* of the specified variable onto the *conditioning variables* (regressors). The first part of the representation of x_t looks just like the $\mathrm{MA}(\infty)$ with *square integrable* moving average terms, while the second part, η_t is something new. That part is called the (linearly) *deterministic part* of x_t because η_t is perfectly *predictable* based on *past* observations of x_t.

The style of proof is constructive, aided by applications of linear projection, and orthogonality and recursive properties of projections. One first finds the d_j 's and ε_t and establish the required properties. Then the projection errors, η_t are found, and its properties are established.

Appendix: The Wold Theorem

Preliminary Results - We begin with a preliminary result. Let x_t be a covariance stationary process. Let

$$\hat{x}_t^{(n)} = P[x_t \mid x_{t-1}, x_{t-2}, \ldots, x_{t-n}] \qquad \text{(A3)}$$

and write

$$x_t = \hat{x}_t^{(n)} + \varepsilon_t^{(n)} \qquad \text{(A4)}$$

From the *orthogonality* property of the projections we know that $\varepsilon_t^{(n)}$ is orthogonal to $(x_{t-1}, x_{t-2}, \ldots, x_{t-n})$ and that $E\,\varepsilon_t^{(n)^2} = \sigma_\varepsilon^{(n)2}$. It can be shown that:

$$\hat{x}_t^{(n)} \to \hat{x} = P[x_t \mid x_{t-1}, x_{t-2}, \ldots] \qquad \text{(A5)}$$

$$x_t = \hat{x}_t + \varepsilon_t \,, \quad E\,\varepsilon_t^2 = \sigma_\varepsilon^2 \,, \quad \varepsilon_t \text{ orthogonal to } (x_{t-1}, x_{t-2}, \ldots)$$

Appendix: The Wold Theorem

The disturbance, ε_t, is known as the 'innovation' in the x_t or its 'one-step-ahead forecast error'. It is easy to see that ε_t is a *serially uncorrelated* process. In particular,

$$\varepsilon_t = x_t - P[x_t \mid x_{t-1}, x_{t-2}, \ldots] \tag{A6}$$

so that it is a linear combination of current and past x_t's. It follows that since ε_t is orthogonal to past x_t's, it is also orthogonal to past ε_t's.

Appendix: The Wold Theorem

Projections of x_t onto Current and Past ε_t's – We now consider the projection of x_t on current and past ε_t s:

$$\tilde{x}_t^{(m)} = \sum_{j=0}^{m} d_j \varepsilon_{t-j} \qquad \text{(A7)}$$

The notation, $\tilde{x}_t^{(m)}$, is intended to signal that the projection used here is *different* from the one used to define ε_t. The lack of autocorrelation between the ε_t's makes the analysis of the projection coefficients particularly simple. The *orthogonality condition* associated with the projection is:

$$E\left(x_t - \sum_{j=0}^{m} d_j \varepsilon_{t-j}\right)\varepsilon_{t-k} = 0 \quad \text{for} \quad k = 0,1,\ldots,m \qquad \text{(A8)}$$

Appendix: The Wold Theorem

which, by the lack of correlation in the ε_t's reduces to:

$$E\,x_t\varepsilon_{t-k} - d_k E\,\varepsilon_{t-k}^2 = 0 \qquad \text{(A9)}$$

so that

$$d_k = \begin{cases} \dfrac{E\,x_t\varepsilon_{t-k}}{\sigma^2}, & k = 1,2,\ldots,m \\ 1, & k = 0 \end{cases} \qquad \text{(A10)}$$

That $E\,x_t\varepsilon_t = \sigma^2$ follows from

$$E\,x_t\varepsilon_t = E(\hat{x}_t + \varepsilon_t)\varepsilon_t = \sigma^2 \qquad \text{(A11)}$$

because $\hat{x}_t$ is a linear function of *past* x_t and ε_t is orthogonal to those x_t's. A key property of the projection is that d_k is *not* a function of *m*. This reflects the lack of serial correlation in the ε_t's.

Appendix: The Wold Theorem

We can establish the *square summability* of the d_j's by noting that any variance must be non-negative and this is true of the error in the projection of x_t onto $\varepsilon_t, \ldots, \varepsilon_{t-m}$:

$$E\left(x_t - \sum_{j=0}^{m} d_j \varepsilon_{t-j}\right)^2 \geq 0 \qquad \text{(A12)}$$

or,

$$E\,x_t^2 - 2\sum_{j=0}^{m} d_j E\,x_t \varepsilon_{t-j} + \sum_{j=0}^{m} d_j^2 \sigma_\varepsilon^2$$

$$= E\,x_t^2 - \sigma_\varepsilon^2 \sum_{j=0}^{m} d_j^2 \geq 0 \qquad \text{(A13)}$$

This must be true for all *m*. Since $E\,x_t^2$ is a *fixed* number by covariance stationarity, it follows that

$$\lim_{m\to\infty} \sum_{j=0}^{m} d_j^2 < \infty \qquad \text{(A14)}$$

Appendix: The Wold Theorem

In addition to the sum is a *non-decreasing* sequence because each term (being square) is non-negative. From this we conclude that the above sum converges to some finite number:

$$\sum_{j=0}^{m} d_j^2 \rightarrow \sum_{j=0}^{\infty} d_j^2 < \infty \qquad \text{(A15)}$$

Given the square summability of the d_j's, it follows that $\tilde{x}_t^{(m)}$ forms a *Cauchy* sequence so that

$$\tilde{x}_t^{(m)} = \sum_{j=0}^{m} d_j \varepsilon_{t-j} \rightarrow \tilde{x}_t = \sum_{j=0}^{\infty} d_j \varepsilon_{t-j} \qquad \text{(A16)}$$

To verify that $\tilde{x}_t^{(m)}$ is a in fact a Cauchy sequence, we establish that for each $\varsigma > 0$, there exists an n such that for $m > n$

$$E\left(\tilde{x}_t^{(m)} - \tilde{x}_t^{(n)}\right)^2 = \left(\sum_{j=n+1}^{m} d_j^2\right)\sigma_\varepsilon^2 < \varsigma \qquad \text{(A17)}$$

Appendix: The Wold Theorem

Constructing the η_t's – We define η_t as the difference between x_t and its projection onto the current and past ε_t's:

$$\eta_t = x_t - \widetilde{x}_t \qquad \text{(A18)}$$

It is relative simple to establish that $E\eta_t\varepsilon_s = 0$ for all *t*, *s*. That $E\eta_t\varepsilon_s = 0$ is true for $s > t$ is obvious because η_t is a linear function of x_t's and past ε_t's:, and ε_s is orthogonal to all these things for $s > t$. That $E\eta_t\varepsilon_s = 0$ for $s \leq t$ follows from the fact that η_t is the error in the projection of x_t on current and past ε_t's. In particular,

$$E\eta_t\varepsilon_{t-k} = Ex_t\varepsilon_{t-k} - d_k\sigma_\varepsilon^2 = d_k\sigma_\varepsilon^2 - d_k\sigma_\varepsilon^2 = 0 \qquad \text{(A19)}$$

What remains is to establish that η_t is *perfectly predictable* from the past x_t's. This proof is somewhat lengthy and is left to the interested reader.

Appendix: The Wold Theorem

Discussion – The Wold representation easy that a covariance stationary process can be represented in the following form:

$$x_t = \underbrace{\sum_{j=0}^{\infty} d_j \varepsilon_{t-j}}_{\text{part of xt that is impossible to predict perfectly}} + \underbrace{\eta_t}_{\text{part of xt that is perfectly predictable}} \quad \text{(A20)}$$

The two parts of this representation are called the 'purely indeterministic' and 'deterministic' parts, respectively. It is interesting to evaluate the meaning of η_t. It is not a time trend, for example, because the assumption of covariance stationarity of x_t rules out a time trend. Here is an example of what η_t could be:

$$\eta_t = a\cos(\lambda t) + b\sin(\lambda t) \quad \text{(A21)}$$

where λ is a fixed number and a and b are (random) variables with

$$Ea = Eb = Eab = 0 \text{ and } a, b \text{ orthogonal to } \{\varepsilon_t\} \quad \text{(A22)}$$

Appendix: The Wold Theorem

To understand this stochastic process for x_t, think of how each realization is constructed. First *draw a* and *b*. Then draw an infinite sequence of ε_t's and generate a realization of η_t and x_t. For the second realization, draw a new *a* and *b*, and a new sequence of ε_t. In this way, all the realizations of stochastic process may be drawn. Under this representation, the mean and autocovariance function of x_t are not a function of time, and so x_t is covariance stationary.

The idea that η_t is perfectly predictable can be seen as follows. First, *a* and *b* can be recovered given only *two observations* on η_t. Once *a* and *b* for a given realization of the stochastic process are in hand, all the η_t in that realization can be computed. But, how to get the two η_t's? According to the argument in the proof, η_t can be *recovered* without error from a suitable linear combination of the $x_{t-1}, x_{t-2}, \ldots$ and η_{t+1} can be recovered from a suitable combination of $x_t, x_{t-1}, \ldots$.

Appendix: The Wold Theorem

Comparison with $MA(\infty)$ Representations - It is interesting to *compare* the purely indeterministic part of the Wold representation with the $MA(\infty)$ representations. The models with $MA(\infty)$ representations are in their most general form, ARMA(p,q) representations:

$$x_t = \varphi_1 x_{t-1} + \ldots + \varphi_p x_{t-p} + \nu_t + \theta_1 \nu_{t-1} + \ldots + \theta_q \nu_{t-q} \tag{A23}$$

where ν_t is i.i.d process. As long as roots of the autoregressive part of this process are less than unity in absolute value, x_t has an $MA(\infty)$ representation with square summable moving average terms. Still, there are two possible *differences* between this and a Wold representation. First, only if the roots of the moving average part, i.e. zeros of

$$\lambda^q + \theta_1 \lambda^{q-1} + \ldots + \theta_q \tag{A24}$$

are less than unity in absolute values is ν_t the one-step-ahead forecast error in x_t (only in this case can recursive substitution be done to represent ν_t as a function of current and all past x_t).

Appendix: The Wold Theorem

Second, the ARMA(p,q) form, while it generates $\mathrm{MA}(\infty)$ with square summable weights, it is not the only form that does this. This is perhaps obvious when we observe that the rate of decay of the moving average coefficients in the models we have considered are geometric. This is a faster rate of decay than is required for square summability. For example, with geometric decay absolute and square summabiilty are the same thing. But in general, a process that is square summable is not necessarily absolutely summable.

We can think of the weights on distant past ε_t's of the $\mathrm{MA}(\infty)$ representation as corresponding to the amount of 'memory' in the process. Thus, the ARMA(p,q) models have 'short memory' relative to the entire class of representations envisioned by the Wold representation. There is some evidence of long-memory in economic time seris, and that this warrants investigating classes of time series models different from the ARMA models.