Time series

Class Notes

Manuel Arellano

March 12, 2017

1 Time series as stochastic outcomes

A time series is a sequence of data points $\{w_t\}_{t=1}^T$ observed over time, typically at equally spaced intervals; for example, the quarterly GDP per capita or the daily number of tweets that mention a specific product. We wish to discuss probabilistic models that regard the observed time series as a realization of a probability distribution function $f(w_1, ..., w_T)$. In a random sample observations are independent and identically distributed so that $f(w_1, ..., w_T) = \prod_{t=1}^T f(w_t)$. In a time series, observations near in time tend to be more similar, in which case the independence assumption is not appropriate. Moreover, the level or other features of the series often change over time, and in that case the assumption of identically distributed observations is not appropriate either. Thus, the joint distribution of the data may differ from the product of marginal distributions of each data point

$$f(w_1, ..., w_T) \neq f_1(w_1) \times f_2(w_2) \times ... \times f_T(w_T)$$

and the form of those marginal distributions may change over time. Due to the natural temporal ordering of data, a factorization that will be often useful is

$$f(w_1, ..., w_T) = f_1(w_1) \prod_{t=2}^T f_t(w_t \mid w_{t-1}, ..., w_1).$$

If the distributions $f_1(.)$, $f_2(.)$, ... changed arbitrarily there would be no regularities on which to base their statistical analysis, since we only have one observation on each distribution. Similarly, if the joint distributions of consecutive pairs of observations changed arbitrarily, there would be no regularities on which to base the analysis of their dependence. Thus, modeling the dependence among observations and their evolving pattern are central to time series analysis. The basic building block to facilitate statistical analysis of time series is some stationary form of dependence that preserves the assumption of identically distributed observations. First we will introduce the concept of stationary dependence and later on we will discuss ways of introducing nonstationarities.

Stochastic process A stochastic process is a collection of random variables that are indexed with respect to the elements in a set of indices. The set may be finite or infinite and contain integer or real numbers. Integer numbers may be equidistant or irregularly spaced. In the case of our time series the set is $\{1, 2, 3, ..., T\}$, but usually it is convenient to consider a set of indices t covering all integers from $-\infty$ to $+\infty$. In such case we are dealing with a double-infinite sequence of the form

$$\{w_t\}_{t=-\infty}^{\infty} = \{..., w_{-1}, w_0, w_1, w_2, ..., w_T, w_{T+1}, ...\},$$

which is regarded as a single realization of the stochastic process, and the observed time series is a portion of this realization. Hypothetical repeated realizations of the process could be indexed as:

$$\left\{w_t^{(1)}, w_t^{(2)}, ..., w_t^{(n)}\right\}_{t=-\infty}^{\infty}.$$

2 Stationarity

A process $\{w_t\}$ is (strictly) stationary if the joint distribution of $\{w_{t_1}, w_{t_2}, ..., w_{t_k}\}$ for a given subset of indices $t_1, t_2, ..., t_k$ is equal to the joint distribution of $\{w_{t_1+j}, w_{t_2+j}, ..., w_{t_k+j}\}$ for any $j > 0$. That is, the distribution of a collection of time points only depends on how far apart they are, and not where they start:

$$f(w_{t_1}, w_{t_2}, ..., w_{t_k}) = f(w_{t_1+j}, w_{t_2+j}, ..., w_{t_k+j}).$$

This implies that marginal distributions are all equal, that the joint distribution of any pair of variables only depends on the time interval between them, and so on:

$$f_{w_t}(.) = f_{w_s}(.) \qquad \text{for all } t, s$$

$$f_{w_t,w_s}(.,.) = f_{w_{t+j},w_{s+j}}(.,.) \qquad \text{for all } t, s, j.$$

In terms of moments, the implication is that the unconditional mean and variance μ_t, σ_t^2 of the distribution $f_{w_t}(.)$ are constant:

$$E(w_t) \equiv \mu_t = \mu, \qquad Var(w_t) \equiv \sigma_t^2 = \sigma^2.$$

Moreover, the covariance $\gamma_{t,s}$ between w_t and w_s only depends on $|t-s|$:

$$Cov(w_t, w_s) \equiv \gamma_{t,s} = \gamma_{|t-s|}.$$

Thus, using the notation $\gamma_0 = \sigma^2$, the covariance matrix of $w = (w_1, ..., w_T)$ takes the form

$$Var(w) = \begin{pmatrix} \gamma_0 & \gamma_1 & & \cdots & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{T-2} \\ \vdots & \gamma_1 & \gamma_0 & \ddots & \vdots \\ \gamma_{T-2} & & & \ddots & \gamma_1 \\ \gamma_{T-1} & \gamma_{T-2} & \cdots & & \gamma_0 \end{pmatrix}.$$

Similarly, the correlation $\rho_{t,s}$ between w_t and w_s only depends on $|t-s|$:

$$\rho_{t,s} \equiv \frac{\gamma_{t,s}}{\sigma_t \sigma_s} = \rho_{|t-s|}$$

The quantity ρ_j is called the autocorrelation of order j and when seen as a function of j it is called the autocorrelation function.

A stationary process is also called strictly stationary in contrast with weaker forms of stationarity. For example, we talk of stationarity in mean if $\mu_t = \mu$ or of covariance stationarity (or weak stationarity) if the process is stationary in mean, variance and covariances. In a normal process covariance stationarity is equivalent to strict stationarity.

Processes that are uncorrelated or independent A sequence of serially uncorrelated random variables with zero mean and constant finite variance is called a "white noise" process; that is, white noise is a covariance stationary process w_t such that

$$\begin{aligned} E(w_t) &= 0 \\ Var(w_t) &= \gamma_0 < \infty \\ Cov(w_t, w_{t-j}) &= 0 \qquad \text{for all } j \neq 0. \end{aligned}$$

In this process observations are uncorrelated but not necessarily independent. In an independent white noise process w_t is also statistically independent of past observations:

$$f(w_t \mid w_{t-1}, w_{t-2}, ...w_1) = f(w_t).$$

Another possibility is a mean independent white noise process that satisfies

$$E(w_t \mid w_{t-1}, w_{t-2}, ...w_1) = 0.$$

In this case w_t is called a martingale difference sequence. A martingale difference is a stronger form of independence than uncorrelatedness but weaker than statistical independence. For example, a martingale difference does not rule out the possibility that $E\left(w_t^2 \mid w_{t-1}, ...w_1\right)$ depends on past observations.

Prediction Consider the problem of selecting a predictor of w_t given a set of past values $\{w_{t-1}, ..., w_{t-j}\}$. The conditional mean $E(w_t \mid w_{t-1}, ..., w_{t-j})$ is the best predictor when the loss function is quadratic. Similarly, the linear projection $E^*(w_t \mid w_{t-1}, ..., w_{t-j})$ is the best linear predictor under quadratic loss. For example, for a stationary process w_t

$$E^*(w_t \mid w_{t-1}) = \alpha + \beta w_{t-1}$$

with $\beta = \gamma_1/\gamma_0$ and $\alpha = (1-\mu)\beta$. We can also write

$$w_t = \alpha + \beta w_{t-1} + \nu_t$$

where ν_t is the prediction error, which by construction is orthogonal to w_{t-1}.

For convenience, predictors based on all past history are often considered:

$$E_{t-1}(w_t) = E(w_t \mid w_{t-1}, w_{t-2}, ...)$$

or

$$E^*_{t-1}(w_t) = E^*(w_t \mid w_{t-1}, w_{t-2}, ...),$$

which are defined as the corresponding quadratic-mean limits of predictors given $\{w_{t-1}, ..., w_{t-j}\}$ as $j \to \infty$.

Linear predictor k-period-ahead Let w_t be a stationary time series with zero mean and let u_t denote the innovation in w_t so that

$$w_t = E^*_{t-1}(w_t) + u_t.$$

u_t is a one-step-ahead forecast error that is orthogonal to all past values of the series. Similarly,

$$w_{t+1} = E^*_t(w_{t+1}) + u_{t+1}.$$

Moreover, since the spaces spanned by $(w_t, w_{t-1}, w_{t-2}, ...)$ and $(u_t, w_{t-1}, w_{t-2}, ...)$ are the same, and u_t is orthogonal to $(w_{t-1}, w_{t-2}, ...)$ we have:

$$E^*_t(w_{t+1}) = E^*(w_{t+1} \mid u_t, w_{t-1}, w_{t-2}, ...) = E^*_{t-1}(w_{t+1}) + E^*(w_{t+1} \mid u_t).$$

Thus, $E^*(w_{t+1} \mid u_t) + u_{t+1}$ is the two-step-ahead forecast error in w_{t+1}. In a similar way we can obtain incremental errors for $E^*_{t-1}(w_{t+2}), ..., E^*_{t-1}(w_{t+k})$.

Wold decomposition Letting $E^*(w_{t+1} \mid u_t) = \psi_1 u_t$, we can write

$$w_{t+1} = u_{t+1} + \psi_1 u_t + E^*_{t-1}(w_{t+1})$$

and repeating the argument we obtain the following representation of the process:

$$w_t = (u_t + \psi_1 u_{t-1} + \psi_2 u_{t-2} + ...) + \kappa_t$$

where $u_t \equiv w_t - E^*_{t-1}(w_t)$, $u_{t-1} \equiv w_{t-1} - E^*_{t-2}(w_{t-1})$, etc. and κ_t denotes the linear prediction of w_t at the beginning of the process. This representation is called the Wold decomposition, after the work of Herman Wold. It exists for any covariance stationary process with zero mean. The one-step-ahead forecast errors u_t are white noise and it can be shown that $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ (with $\psi_0 = 1$).[1]

The term κ_t is called the linearly deterministic part of w_t because it is perfectly predictable based on past observations of w_t. The other part, consisting of $\sum_{j=0}^{\infty} \psi_j u_{t-j}$, is the linearly indeterministic part of the process. The indeterministic part is the linear projection of w_t onto the current and past linear forecast errors, and the deterministic part is the corresponding projection error. If $\kappa_t = 0$, w_t is a purely non-deterministic process, also called a linearly regular process.

[1] See T. Sargent, *Macroeconomic Theory*, 1979.

Ergodicity A stochastic process is ergodic if it has the same behavior averaged over time as averaged over the sample space. Specifically, a covariance stationary process is ergodic in mean if the time series mean converges in probability to the same limit as a (hypothetical) cross-sectional mean (known as the ensemble average), that is, to $E(w_t) = \mu$:

$$\overline{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t \xrightarrow{p} \mu.$$

Ergodicity requires that the autocovariances γ_j tend to zero sufficiently fast as j increases. In the next section we check that $\{w_t\}$ is ergodic in mean if the following absolute summability condition is satisfied:

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty. \tag{1}$$

Similarly, a covariance stationary process is ergodic in covariance if

$$\frac{1}{T-j}\sum_{t=j+1}^{T} (w_t - \mu)(w_{t-j} - \mu) \xrightarrow{p} \gamma_j.$$

In the special case in which $\{w_t\}$ is a normal stationary process, condition (1) guarantees ergodicity for all moments.[2]

Example of stationary non-ergodic process Suppose that

$$w_t = \eta + \varepsilon_t$$

where $\eta \sim iid\left(0, \sigma_\eta^2\right)$ and $\varepsilon_t \sim iid\left(0, \sigma_\varepsilon^2\right)$ independent of η. We have

$$\begin{aligned}
\mu &= E(w_t) = E(\eta) + E(\varepsilon_t) = 0 \\
\gamma_0 &= Var(w_t) = Var(\eta) + Var(\varepsilon_t) = \sigma_\eta^2 + \sigma_\varepsilon^2 \\
\gamma_j &= Cov(w_t, w_{t-j}) = \sigma_\eta^2 \qquad \text{for } j \neq 0.
\end{aligned}$$

Note that condition (1) is not satisfied in this example.

Let the index i denote a realization of the process in the probability space. The process is stationary and yet

$$\overline{w}_T^{(i)} = \frac{1}{T}\sum_{t=1}^{T} w_t^{(i)} \xrightarrow{p} \eta^{(i)}$$

instead of converging to $\mu = 0$. Moreover,

$$\frac{1}{T-j}\sum_{t=j+1}^{T} \left(w_t^{(i)} - \eta^{(i)}\right)\left(w_{t-j}^{(i)} - \eta^{(i)}\right) \xrightarrow{p} 0$$

(or $(T-j)^{-1}\sum_{t=j+1}^{T} w_t^{(i)} w_{t-j}^{(i)} \xrightarrow{p} \left(\eta^{(i)}\right)^2$) instead of converging to $\gamma_j = \sigma_\eta^2$.

[2] In general, a stationary process is ergodic in distribution if $T^{-1}\sum_{t=1}^{T} \mathbf{1}(w_t \leq r) \xrightarrow{p} \Pr(w_t \leq r)$ for any r, where $\mathbf{1}(w_t \leq r) = 1$ if $w_t \leq r$ and $\mathbf{1}(w_t \leq r) = 0$ if $w_t > r$.

3 Asymptotic theory with dependent observations

Here we consider a law of large numbers and a central limit theorem for covariance stationary processes.

Law of Large Numbers Let $\{w_t\}$ be a covariance stationary stochastic process with $E(w_t) = \mu$ and $Cov(w_t, w_{t-j}) = \gamma_j$ such that $\sum_{j=0}^{\infty} |\gamma_j| < \infty$. Let the sample mean be $\overline{w}_T = (1/T)\sum_{t=1}^{T} w_t$. Then (i) $\overline{w}_T \xrightarrow{p} \mu$, and (ii) $Var\left(\sqrt{T}\overline{w}_T\right) \to \sum_{j=-\infty}^{\infty} \gamma_j$.

A sufficient condition for $\overline{w}_T \xrightarrow{p} \mu$ is that $E(\overline{w}_T) \to \mu$ and $Var(\overline{w}_T) \to 0$. For any T we have $E(\overline{w}_T) = \mu$. Next,

$$\begin{aligned} Var(\overline{w}_T) &= E\left[(\overline{w}_T - \mu)^2\right] = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} E\left[(w_t - \mu)(w_s - \mu)\right] \\ &= \frac{1}{T^2}\left[T\gamma_0 + 2(T-1)\gamma_1 + 2(T-2)\gamma_2 + ... + 2\gamma_{T-1}\right]. \end{aligned}$$

To show that $Var(\overline{w}_T) \to 0$, show that $TVar(\overline{w}_T)$ is bounded under the assumption $\sum_{j=0}^{\infty} |\gamma_j| < \infty$:

$$\begin{aligned} TVar(\overline{w}_T) &= \left|\gamma_0 + \left(\frac{T-1}{T}\right)2\gamma_1 + \left(\frac{T-2}{T}\right)2\gamma_2 + ... + \frac{1}{T}2\gamma_{T-1}\right| \\ &\leq |\gamma_0| + \left(\frac{T-1}{T}\right)2|\gamma_1| + \left(\frac{T-2}{T}\right)2|\gamma_2| + ... + \frac{1}{T}2|\gamma_{T-1}| \\ &\leq \left\{|\gamma_0| + 2|\gamma_1| + 2|\gamma_2| + ... + 2|\gamma_{T-1}| + ...\right\}. \end{aligned} \tag{2}$$

To check that $Var\left(\sqrt{T}\overline{w}_T\right) \to \sum_{j=-\infty}^{\infty} \gamma_j$ see J. Hamilton, *Time Series Analysis*, 1994, p. 187–188.

Consistent estimation of second-order moments Let us consider the sample autocovariance

$$\widehat{\gamma}_j = \frac{1}{T-j}\sum_{t=j+1}^{T}(w_t - \overline{w}_0)(w_{t-j} - \overline{w}_{-j}) = \frac{1}{T-j}\sum_{t=j+1}^{T} w_t w_{t-j} - \overline{w}_0\overline{w}_{-j}$$

where $\overline{w}_0 = (T-j)^{-1}\sum_{t=j+1}^{T} w_t$ and $\overline{w}_{-j} = (T-j)^{-1}\sum_{t=j+1}^{T} w_{t-j}$. Let us define $z_t = w_t w_{t-j}$. The previous LLN can be applied to the process z_t to state conditions under which

$$\frac{1}{T-j}\sum_{t=j+1}^{T} w_t w_{t-j} \xrightarrow{p} E(w_t w_{t-j}).$$

Note that if w_t is strictly stationary so is z_t.

Central Limit Theorem A central limit theorem provides conditions under which

$$\frac{\overline{w}_T - \mu}{\sqrt{Var(\overline{w}_T)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Since in our context $TVar(\overline{w}_T) \to \sum_{j=-\infty}^{\infty} \gamma_j$, an asymptotically equivalent statement is

$$\frac{\sqrt{T}(\overline{w}_T - \mu)}{\sqrt{\sum_{j=-\infty}^{\infty} \gamma_j}} \xrightarrow{d} \mathcal{N}(0,1).$$

or

$$\sqrt{T}\left(\overline{w}_T - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=-\infty}^{\infty} \gamma_j\right). \tag{3}$$

A condition under which this result holds is:

$$w_t = \mu + \sum_{j=0}^{\infty} \psi_j v_{t-j} \tag{4}$$

where $\{v_t\}$ is an i.i.d. sequence with $E\left(v_t^2\right) < \infty$ and $\sum_{j=0}^{\infty} \left|\psi_j\right| < \infty$ (T.W. Anderson, *The Statistical Analysis of Time Series*, 1971, p. 429). Result (3) also holds when the innovation process $\{v_t\}$ in (4) is a martingale difference sequence satisfying certain conditions.[3]

A multivariate version of (3) for the case in which $\{w_t\}$ is a vector-valued process is as follows:

$$\sqrt{T}\left(\overline{w}_T - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=-\infty}^{\infty} \Gamma_j\right). \tag{5}$$

where

$$\Gamma_0 = E\left[\left(w_t - \mu\right)\left(w_t - \mu\right)'\right]$$

and for $j \neq 0$:

$$\Gamma_j = E\left[\left(w_t - \mu\right)\left(w_{t-j} - \mu\right)'\right].$$

Note that the autocovariance matrix Γ_j is not symmetric. We have $\Gamma_{-j} = \Gamma_j'$.

As in the scalar case, a condition under which result (5) holds is:

$$w_t = \mu + \sum_{j=0}^{\infty} \Psi_j v_{t-j}$$

where $\{v_t\}$ is an i.i.d. vector sequence with $E\left(v_t\right) = 0$, $E\left(v_t v_t'\right) = \Omega$ a symmetric positive definite matrix, and the sequence of matrices $\{\Psi_j\}_{j=0}^{\infty}$ is absolutely summable.[4]

Consistent estimation of the asymptotic variance To be able to use the previous central limit theory for the construction of interval estimations and test statistics we need consistent estimators of $V = \sum_{j=-\infty}^{\infty} \gamma_j$. One possibility is to parameterize γ_j; for example, assuming that the γ_j satisfy the restrictions imposed by an ARMA model of the type that are discussed in the next section. Another possibility is to obtain a flexible estimator of V of the type considered by Hansen (1982), and Newey and West (1987), among others.[5]

The Newey-West estimator is a sample counterpart of expression (2) truncated after m lags:

$$\widehat{V} = \widehat{\gamma}_0 + \sum_{j=1}^{m} \left(1 - \frac{j}{m+1}\right) 2\widehat{\gamma}_j.$$

[3] Theorem 3.15 in P.C.B. Phillips and V. Solo (1992) "Asymptotics for Linear Processes," *The Annals of Statistics* 20.

[4] The matrix sequence $\{\Psi_j\}_{j=0}^{\infty}$ is absolutely summable if each of its elements forms an absolutely summable sequence.

[5] Hansen (1982) "Large Sample Properties of GMM Estimators", *Econometrica* 50. Newey and West (1987) "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica* 55.

This estimator can be shown to be consistent for V if the truncation parameter m goes to infinity with T more slowly than $T^{1/4}$ or than $T^{1/2}$, depending on the type of process. Nevertheless, the specification of an appropriate growth rate for m gives little practical guidance on the choice of m.

Similarly, in the vector case the Newey-West estimator of $V = \sum_{j=-\infty}^{\infty} \Gamma_j$ is given by

$$\widehat{V} = \widehat{\Gamma}_0 + \sum_{j=1}^{m} \left(1 - \frac{j}{m+1}\right) \left(\widehat{\Gamma}_j + \widehat{\Gamma}_j'\right) \tag{6}$$

where $\widehat{\Gamma}_j = (T-j)^{-1} \sum_{t=j+1}^{T} (w_t - \overline{w}_0)(w_{t-j} - \overline{w}_{-j})'$. A nice property of the Newey-West estimator (6) is that it is guaranteed to be a positive semi-definite matrix by construction.[6]

4 Autoregressive and moving average models

4.1 Autoregressive models

A first-order autoregressive process (or Markov process) assumes that w_t is independent of $\{w_{t-2}, w_{t-3}, ...\}$ conditionally on w_{t-1}:

$$f_t(w_t \mid w_{t-1}, ..., w_1) = f_t(w_t \mid w_{t-1})$$

and, therefore, also

$$E(w_t \mid w_{t-1}, ..., w_1) = E(w_t \mid w_{t-1}).$$

Moreover, the standard linear AR(1) model also assumes

$$\begin{aligned} E(w_t \mid w_{t-1}) &= \alpha + \rho w_{t-1} \\ Var(w_t \mid w_{t-1}) &= \sigma^2. \end{aligned}$$

Moment properties These assumptions have the following implications for marginal moments:

$$E(w_t) = E[E(w_t \mid w_{t-1})] = \alpha + \rho E(w_{t-1}) \tag{7}$$

$$Var(w_t) = Var[E(w_t \mid w_{t-1})] + E[Var(w_t \mid w_{t-1})] = \rho^2 Var(w_{t-1}) + \sigma^2$$

$$Cov(w_t, w_{t-1}) = Cov(E(w_t \mid w_{t-1}), w_{t-1}) = Cov(\alpha + \rho w_{t-1}, w_{t-1}) = \rho Var(w_{t-1}).$$

Moreover,

$$E(w_t \mid w_{t-2}) = \alpha + \rho E(w_{t-1} \mid w_{t-2}) = \alpha(1+\rho) + \rho^2 w_{t-2}.$$

[6] The estimator $\widetilde{V} = \widehat{\Gamma}_0 + \sum_{j=1}^{m} \left(\widehat{\Gamma}_j + \widehat{\Gamma}_j'\right)$ has the same large sample justification as (6) but is not necessarily positive semi-definite.

In general

$$E\left(w_t \mid w_{t-j}\right) = \alpha\left(1+\rho+...+\rho^{j-1}\right)+\rho^j w_{t-j}$$

and

$$Cov\left(w_t, w_{t-j}\right) = Cov\left(E\left(w_t \mid w_{t-j}\right), w_{t-1}\right) = \rho^j Var\left(w_{t-j}\right).$$

In view of the recursion (7) we have

$$E\left(w_t\right) = \alpha\left(1+\rho+...+\rho^{t-1}\right)+\rho^t E\left(w_0\right).$$

Stability and stationarity For the process to be stationary is required that $|\rho| < 1$. In itself, $|\rho| < 1$ is a condition of stability under which as $t \to \infty$ we obtain

$$E\left(w_t\right) \to \mu = \frac{\alpha}{1-\rho}$$

$$Var\left(w_t\right) \to \gamma_0 = \frac{\sigma^2}{1-\rho^2}$$

$$Cov\left(w_t, w_{t-j}\right) \to \gamma_j = \rho^j \frac{\sigma^2}{1-\rho^2}.$$

These quantities are known as the steady state mean, variance and autocovariances. Thus, regardless of the starting point, under the stability condition the AR(1) process is asymptotically covariance stationary.

If the AR(1) process is stationary (due to being stable and having started in the distant past or having started at $t = 1$ with the steady state distribution) then $E\left(w_t\right) = \mu = \alpha/\left(1-\rho\right)$, $Var\left(w_t\right) = \gamma_0 = \sigma^2/\left(1-\rho^2\right)$ and $Cov\left(w_t, w_{t-j}\right) = \gamma_j = \rho^j\sigma^2/\left(1-\rho^2\right)$.

The autocorrelation function of a stationary AR(1) process decreases exponentially and is given by ρ^j.

Letting $u_t = w_t - \alpha - \rho w_{t-1}$, the Wold representation of a stationary AR(1) process can be obtained by repeated substitutions and is given by:

$$w_t = \mu + u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + ...$$

The parameter ρ measures the persistence in the process. The closer is ρ to one the more persistent will be the deviations of the process from its mean.

Normality assumptions The standard additional assumption to fully specify the distribution of $w_t \mid w_{t-1}$ is conditional normality:

$$w_t \mid w_{t-1}, ..., w_1 \sim \mathcal{N}\left(\alpha+\rho w_{t-1}, \sigma^2\right). \tag{8}$$

In itself this assumption does not imply unconditional normality. However, if we assume that the initial observation is normally distributed with the steady state mean and variance:

$$w_1 \sim \mathcal{N}\left(\frac{\alpha}{1-\rho}, \frac{\sigma^2}{1-\rho^2}\right), \tag{9}$$

then the process is fully stationary and $(w_1, ..., w_T)$ is jointly normally distributed as follows:

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_T \end{pmatrix} \sim \mathcal{N}\left[\frac{\alpha}{1-\rho}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \frac{\sigma^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \dots & \rho^{T-1} \\ \rho & 1 & \dots & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \dots & 1 \end{pmatrix}\right].$$

Normal likelihood functions Under assumption (8), the log likelihood function of the time series $\{w_1, ..., w_T\}$ conditioned on the first observation is given by (up to an additive constant):

$$L\left(\alpha, \rho, \sigma^2\right) = -\frac{(T-1)}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(w_t - \alpha - \rho w_{t-1}\right)^2.$$

The corresponding maximum likelihood estimates are:

$$\begin{aligned} \widehat{\rho} &= \frac{\sum_{t=2}^{T}\left(w_t - \overline{w}_0\right)\left(w_{t-1} - \overline{w}_{-1}\right)}{\sum_{t=2}^{T}\left(w_{t-1} - \overline{w}_{-1}\right)^2} \\ \widehat{\alpha} &= \overline{w}_0 - \widehat{\rho}\overline{w}_{-1} \\ \widehat{\sigma}^2 &= \frac{1}{(T-1)}\sum_{t=2}^{T}\left(w_t - \widehat{\alpha} - \widehat{\rho} w_{t-1}\right)^2 \end{aligned} \tag{10}$$

where $\overline{w}_0 = (T-1)^{-1}\sum_{t=2}^{T} w_t$ and $\overline{w}_{-1} = (T-1)^{-1}\sum_{t=2}^{T} w_{t-1}$.

Under the steady state assumption (9) the log likelihood of the first observation is given by:

$$\ell_1\left(\alpha, \rho, \sigma^2\right) = -\frac{1}{2}\ln\sigma^2 + \frac{1}{2}\ln\left(1-\rho^2\right) - \frac{\left(1-\rho^2\right)}{2\sigma^2}\left(w_1 - \frac{\alpha}{1-\rho}\right)^2.$$

Thus, the full log likelihood function under assumptions (8) and (9) becomes:

$$L^*\left(\alpha, \rho, \sigma^2\right) = L\left(\alpha, \rho, \sigma^2\right) + \ell_1\left(\alpha, \rho, \sigma^2\right).$$

The estimators that maximize $L^*\left(\alpha, \rho, \sigma^2\right)$ lack a closed form expression.

Asymptotic properties of OLS estimates in the AR(1) model Let us focus on the OLS estimate of the autoregressive parameter (10) when $|\rho| < 1$. Since $\widehat{\rho} = \widehat{\gamma}_1/\widehat{\gamma}_0$, consistency of $\widehat{\rho}$ for $\rho = \gamma_1/\gamma_0$ follows under conditions ensuring the consistency of sample autocovariances. Next, consider the scaled estimation error and its large-sample approximation:

$$\begin{aligned} \sqrt{T}\left(\widehat{\rho} - \rho\right) &= \left[\frac{1}{T}\sum_{t=2}^{T}\left(w_{t-1} - \overline{w}_{-1}\right)^2\right]^{-1}\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\left(w_{t-1} - \overline{w}_{-1}\right)u_t \\ &\approx \left(\frac{\sigma^2}{1-\rho^2}\right)^{-1}\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\left(w_{t-1} - \mu\right)u_t. \end{aligned}$$

Under conditions ensuring the asymptotic normality result

$$T^{-1/2}\sum_{t=2}^{T}(w_{t-1}-\mu)\,u_t \xrightarrow{d} \mathcal{N}(0,\omega)$$

with $\omega = E\left[u_t^2\,(w_{t-1}-\mu)^2\right] = \sigma^4/\left(1-\rho^2\right)$, we obtain

$$\sqrt{T}\,(\widehat{\rho}-\rho) \xrightarrow{d} \mathcal{N}\left(0, 1-\rho^2\right).$$

When $\rho \geq 1$ this result does not hold and the OLS properties are non-standard.

Forecasting with a stable AR(1) process A one-period-ahead forecast is

$$E_t(w_{t+1}) = \alpha + \rho w_t,$$

a two-period ahead forecast is

$$E_t(w_{t+2}) = \alpha(1+\rho) + \rho^2 w_t,$$

and k-period ahead

$$E_t(w_{t+k}) = \alpha\left(1+\rho+...+\rho^{k-1}\right) + \rho^k w_t = \left(\frac{\alpha}{1-\rho}\right)\left(1-\rho^k\right) + \rho^k w_t.$$

Thus, a k-period ahead forecast is a convex combination of the steady state mean and the most recent value of the process available. As $k \to \infty$ the optimal forecast tends to the steady state mean $\alpha/(1-\rho)$.

AR(p) process A generalization of the AR(1) process is to an AR(p) process that specifies linear dependence on the first p lags:

$$w_t = \alpha + \rho_1 w_{t-1} + ... + \rho_p w_{t-p} + u_t.$$

Second-order or higher-order autoregressive processes can capture richer patterns of behavior in time series, including stochastic cycles.

4.2 Moving average models

To motivate the moving average model, consider the stationary linear process with iid shocks in (4):

$$w_t = \mu + u_t + \psi_1 u_{t-1} + \psi_2 u_{t-2} + ...$$

The independent white noise process is the special case when $\psi_j = 0$ for all $j \geq 1$ and $\mu = 0$. A first-order moving average process relaxes the independence assumption by allowing ψ_1 to be non-zero while setting $\psi_j = 0$ for $j > 1$. Thus, the form of an MA(1) process is

$$w_t = \mu + u_t - \theta u_{t-1}$$

where $u_t \sim iid\left(0, \sigma^2\right)$.

Moment properties In this case

$$\begin{aligned} E(w_t) &= \mu \\ Var(w_t) &= \gamma_0 = \left(1+\theta^2\right)\sigma^2 \\ Cov(w_t, w_{t-1}) &= \gamma_1 = -\theta\sigma^2 \\ Cov(w_t, w_{t-j}) &= \gamma_j = 0 \text{ for } j > 1. \end{aligned}$$

Note that the MA(1) process is stationary for all values of θ.

The first-order autocorrelation is

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = -\frac{\theta}{1+\theta^2},$$

which means that $-0.5 \leq \rho_1 \leq 0.5$.

Indeterminacy and invertibility The moving average parameter θ solves:

$$\rho_1\theta^2 + \theta + \rho_1 = 0. \tag{11}$$

The product of the roots of this equation is unity,[7] so that if θ is a solution, then $1/\theta$ is also a solution. Moreover, if one solution is less than unity in absolute value, the other one will be greater than unity.

If $|\theta| < 1$ then it can be seen that the MA(1) model can be written as a convergent series of past values of the process:

$$w_t + \sum_{j=1}^{\infty} \theta^j w_{t-j} = u_t.$$

If on the contrary $|\theta| > 1$, the MA(1) model can also be written as a convergent series but one involving the future values of the process:

$$w_t + \sum_{j=1}^{\infty} \frac{1}{\theta^j} w_{t+j} = u_t.$$

Given a preference for associating present values of the process with past values, the indeterminacy about the value of θ is avoided by requiring that $|\theta| < 1$, a condition called "invertibility" by Box and Jenkins (*Time Series Analysis: Forecasting and Control*, 1976).[8]

Normal likelihood function Under joint normality $w = (w_1, ..., w_T)' \sim \mathcal{N}\left[\mu\iota, \sigma^2\Omega(\theta)\right]$ with ι denoting a $T \times 1$ vector of ones and

$$\Omega(\theta) = \begin{pmatrix} 1+\theta^2 & -\theta & \dots & 0 \\ -\theta & 1+\theta^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1+\theta^2 \end{pmatrix},$$

[7] The roots are $\left(-1+\sqrt{1-4\rho_1^2}\right)/(2\rho_1)$ and $\left(-1-\sqrt{1-4\rho_1^2}\right)/(2\rho_1)$ provided $\rho_1 \neq 0$.

[8] If $|\rho_1| = 0.5$ there is a unique non-invertible solution to (11) such that $|\theta| = 1$.

the log likelihood function of the time series is given by

$$L\left(\mu,\theta,\sigma^2\right) = -\frac{T}{2}\ln\sigma^2 - \frac{1}{2}\ln\Omega\left(\theta\right) - \frac{1}{2\sigma^2}\left(w-\mu\iota\right)'\left[\Omega\left(\theta\right)\right]^{-1}\left(w-\mu\iota\right).$$

There is no closed form expression for the maximum likelihood estimator. The natural estimator of μ is the sample mean $\overline{w}_T$. A simple consistent estimator $\widehat{\theta}$ is the invertible solution to the equation:

$$\widehat{\rho}_1\theta^2 + \theta + \widehat{\rho}_1 = 0$$

where $\widehat{\rho}_1 = \widehat{\gamma}_1/\widehat{\gamma}_0$. The corresponding estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{\widehat{\gamma}_0}{1+\widehat{\theta}^2}.$$

Forecasting with an invertible MA(1) process The autoregressive representation of the process is

$$w_t = \left(\frac{1-\theta^t}{1-\theta}\right)\mu - \theta w_{t-1} - ... - \theta^{t-1}w_1 - \theta^t u_0 + u_t$$

and a similar expression one period ahead

$$w_{t+1} = \left(\frac{1-\theta^{t+1}}{1-\theta}\right)\mu - \theta w_t - ... - \theta^t w_1 - \theta^{t+1}u_0 + u_{t+1}.$$

Thus, an infeasible forecast based on all past history is:

$$E_t\left(w_{t+1}\right) = \left(\frac{1-\theta^{t+1}}{1-\theta}\right)\mu - \theta w_t - ... - \theta^t w_1 - \theta^{t+1}E_t\left(u_0\right).$$

An approximate feasible forecast ignores the last term that contains $E_t\left(u_0\right)$. Alternatively, we may calculate the best linear predictor of w_{t+1} given $\{w_1,...,w_t\}$ taking into account that $E\left(w\right) = \mu\iota$ and $Var\left(w\right) = \sigma^2\Omega\left(\theta\right)$. For example,

$$E^*\left(w_{T+1} \mid w_T,...,w_1\right) = \delta + w'\varphi$$

where $\varphi = \left[\Omega\left(\theta\right)\right]^{-1}q\left(\theta\right)$, $q\left(\theta\right) = \left(-\theta,0,...,0\right)$ and $\delta = \mu\left(1-\iota'\varphi\right)$.

MA(q) process A generalization of the MA(1) process is to an MA(q) process that specifies linear dependence on the first q shocks:

$$w_t = \mu + u_t - \theta_1 u_{t-1} - ... - \theta_q u_{t-q}.$$

ARMA (p,q) process A further generalization is a process that combines an autoregressive component and a moving average component. For example, the ARMA(1,1) process takes the form:

$$w_t = \alpha + \rho w_{t-1} + u_t - \theta u_{t-1}.$$

An ARMA process may be able to approximate a linear process to a given accuracy employing fewer parameters than a pure autoregressive or a pure moving average process.

5 Nonstationary processes

5.1 Time trends

In a stationary process $E(w_t) = \mu$. A less restrictive assumption that allows for nonstationarity in mean is to specify the mean as a function of time. For example, a linear trend:

$$E(w_t) = \alpha + \beta t.$$

If w_t represents the logarithm of some variable, β is a growth rate, which in this model is assumed to be constant.

We could assume that

$$w_t = \alpha + \beta t + u_t$$

where u_t is a stationary stochastic process. In such case, $E(w_t) = \alpha + \beta t$ but $Var(w_t)$ is constant.

In the same vein, the specification of the mean of the process could incorporate cyclical or seasonal components.

Regression with trend In a regression with a linear trend, OLS estimation errors converge to zero at a faster rate than the usual root-T consistency. The reason is that the second moment of a conventional regressor is bounded whereas the second moment of a linear trend is not. To examine this situation let us consider a simple linear trend model with an iid normal error and no intercept:

$$y_t = \beta t + u_t \qquad u_t \sim iid\, \mathcal{N}\left(0, \sigma^2\right). \tag{12}$$

The OLS estimation error and the OLS mean and variance are given by

$$\widehat{\beta} - \beta = \frac{\sum_{t=1}^{T} t\, u_t}{\sum_{t=1}^{T} t^2}.$$

$$E\left(\widehat{\beta}\right) = \beta \qquad Var\left(\widehat{\beta}\right) = \frac{\sigma^2}{\sum_{t=1}^{T} t^2} = \frac{\sigma^2}{T(T+1)(2T+1)/6}.$$

This is a classical regression model with $x_t = t$, no intercept, and normal errors, except that in the standard model $\sum_{t=1}^{T} x_t^2 = O(T)$ whereas here $\sum_{t=1}^{T} t^2 = O\left(T^3\right)$.

In this case the following exact distributional result holds:

$$\left(\sum_{t=1}^{T} t^2\right)^{1/2} \frac{\left(\widehat{\beta} - \beta\right)}{\sigma} \sim \mathcal{N}(0,1)$$

and therefore also as $T \to \infty$:

$$\left(\sum_{t=1}^{T} t^2\right)^{1/2} \frac{\left(\widehat{\beta} - \beta\right)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$$

or

$$T^{3/2}\left[\frac{1}{6}\left(1+\frac{1}{T}\right)\left(2+\frac{1}{T}\right)\right]^{1/2}\frac{\left(\widehat{\beta}-\beta\right)}{\sigma}\xrightarrow{d}\mathcal{N}(0,1)$$

and

$$T^{3/2}\left(\frac{1}{6}\times 1\times 2\right)^{1/2}\frac{\left(\widehat{\beta}-\beta\right)}{\sigma}\xrightarrow{d}\mathcal{N}(0,1),$$

and also

$$T^{3/2}\left(\widehat{\beta}-\beta\right)\xrightarrow{d}\mathcal{N}\left(0,3\sigma^2\right). \tag{13}$$

It can be shown that result (13) still holds if $u_t \sim iid\left(0,\sigma^2\right)$ but non-normal.[9] Thus, (13) justifies the large-sample approximation

$$\widehat{\beta}\approx\mathcal{N}\left(\beta,\frac{3\sigma^2}{T^3}\right).$$

This situation is described as $T^{3/2}$-consistency (in contrast with $T^{1/2}$-consistency) and $\widehat{\beta}$ is said to be "hyper-consistent" or "super-consistent" (although the last term sometimes is reserved for T-consistent estimators).

5.2 Random walk

A random walk is a process such that

$$E\left(w_t \mid w_{t-1}, w_{t-2}, ...\right) = w_{t-1},$$

so that the best forecast is the previous value and there is no mean reversion.

A random walk with iid shocks is an AR(1) model with $\rho = 1$:

$$w_t = w_{t-1} + u_t \qquad u_t \sim iid\left(0,\sigma^2\right) \tag{14}$$

or equivalently

$$w_t = u_t + u_{t-1} + ... + u_1 + w_0,$$

and the first-difference $\Delta w_t = (w_t - w_{t-1})$ is an independent white noise process.

The random walk is a nonstationary process. Letting $\omega_0 = Var\left(w_0\right)$, we have

$$\begin{aligned} Var\left(w_t\right) &= \omega_0 + t\sigma^2 \\ Cov\left(w_t, w_{t-j}\right) &= \omega_0 + (t-j)\,\sigma^2 \qquad \text{for } j \geq 1. \end{aligned}$$

[9] See, for example, T.W. Anderson, *The Statistical Analysis of Time Series*, 1971, Theorem 2.6.1, p. 23.

Thus, the variance tends to infinity as t increases and the autocorrelation function decays slowly as j increases.

More generally, we could consider processes such as (14) in which u_t is a stationary process, but not necessarily a white noise process. A time series such that its first difference is a stationary process is called a first-order integrated process or an $I(1)$ process:

$$w_t \sim I(1).$$

If u_t is an ARMA(p,q) process then w_t is called an autoregressive integrated moving average or ARIMA$(p,1,q)$ process. The argument can be generalized to second and higher-order integrated process. For example, an $I(2)$ process is such that it is stationary after taking differences twice.

Random walk with drift This is a process of the form

$$w_t = w_{t-1} + \delta + u_t \qquad u_t \sim iid\left(0, \sigma^2\right).$$

In this case:

$$w_t = \delta t + (u_t + u_{t-1} + ... + u_1 + w_0).$$

We have a linear trend, but contrary to (12) the stochastic component is $I(1)$ instead of $I(0)$.

Distinguishing between unit root and stationary processes There is a literature concerned with large sample methods to test the null hypothesis of a unit root in w_t against the alternative hypothesis of stationarity (e.g. the Dickey-Fuller test and its variants). However, a realization from an $I(1)$ process may be difficult to distinguish from an $I(0)$ process. Compare, for example, the following integrated moving average process

$$w_t - w_{t-1} = u_t - 0.99u_{t-1}$$

with the white noise process

$$w_t = u_t;$$

or the random walk process

$$w_t - w_{t-1} = u_t$$

with the stationary autoregressive process

$$w_t - 0.99w_{t-1} = u_t.$$

The differences between those $I(1)$ and $I(0)$ processes are in their long run behavior. Sometimes the choice between an $I(1)$ model and an $I(0)$ model is made on the basis of the long-run properties that are judged a priori to make sense for a time series to have.