Applied time-series analysis

Robert M. Kunst
robert.kunst@univie.ac.at

University of Vienna
and
Institute for Advanced Studies Vienna

October 18, 2011

Outline

Econometric Time-Series Analysis

In principle, time-series analysis is a field of statistics.
Special features of economic time series:

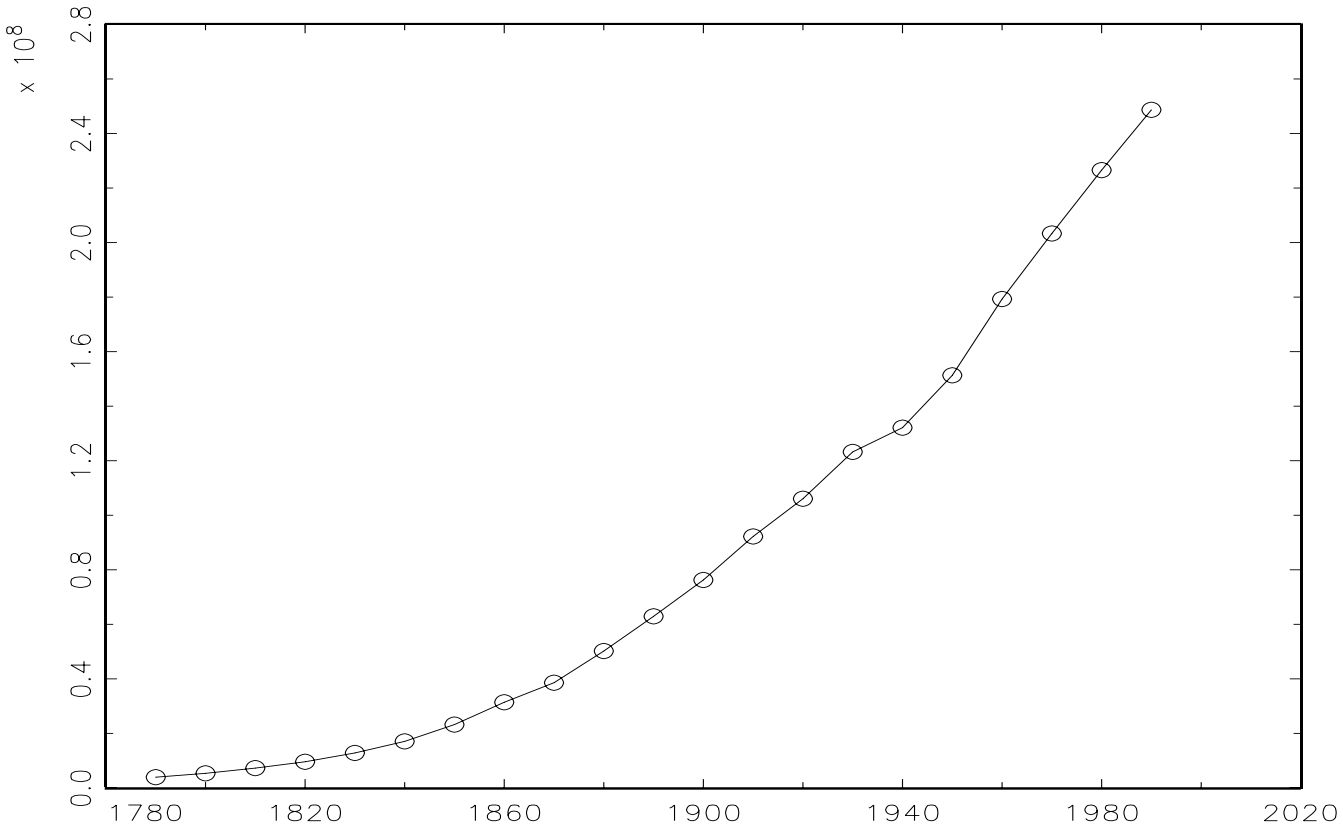
- Economic time series are often short: models must be parsimonious, frequency domain is not attractive.
- Trends: emphasis on difference stationarity ('unit roots') and cointegration.
- Seasonal cycles: seasonal adjustment or seasonal time-series models.
- Economic theory plays a leading role: attempts to integrate data-driven time-series analysis and theory-based structure.
- Finance time series are different: long series, volatility modelling (ARCH), nonlinear models.

The idea of time-series analysis

The observed time series is seen as a realization of a stochastic process. Using the assumption of some degree of time constancy, the data should indicate a potential and reasonable data-generating process (DGP).

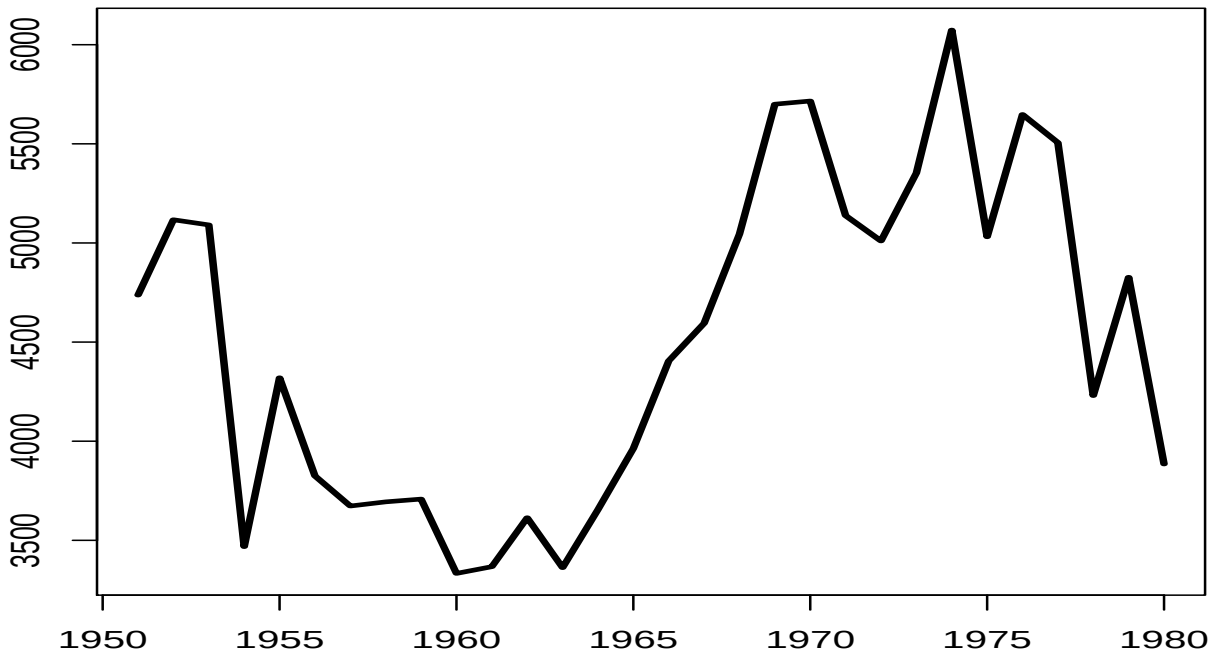
This concept has proven to be more promising than non-stochastic approaches: curve fitting, extrapolation.

Examples of time series: U.S. population



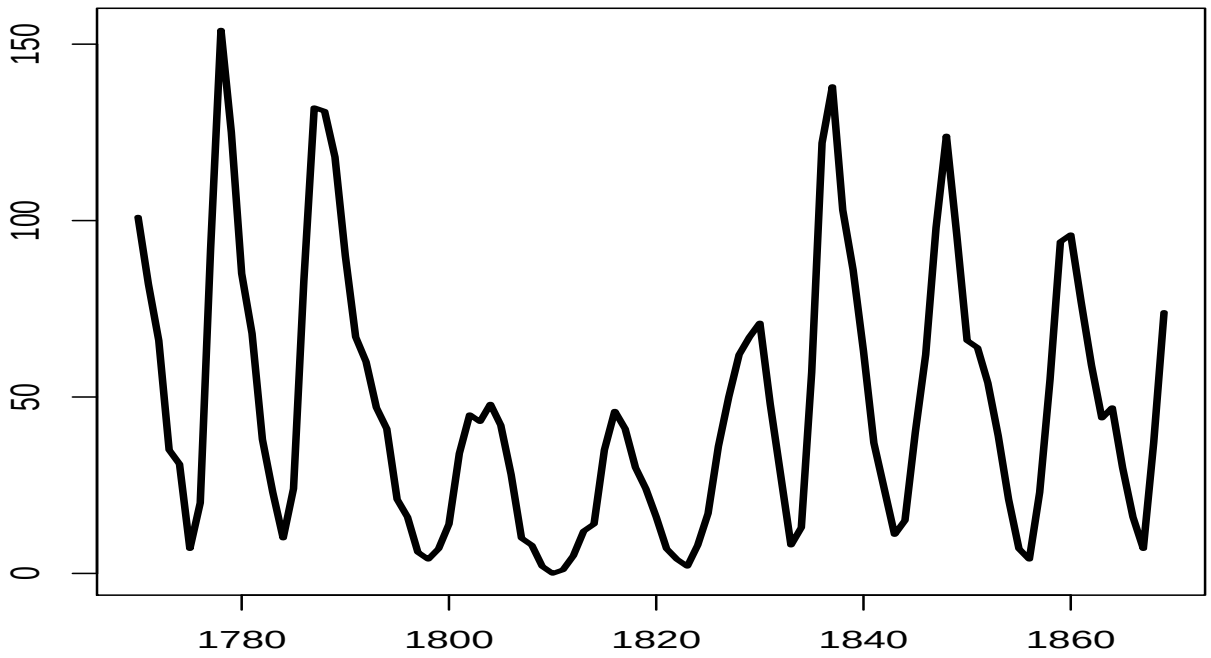
Population measured at ten-year intervals 1790–1990 (from Brockwell and Davis).

Examples of time series: Strikes in the U.S.A.



Strikes in the U.S.A., annual data 1951–1980 (from Brockwell and Davis).

Examples of time series: Sunspots

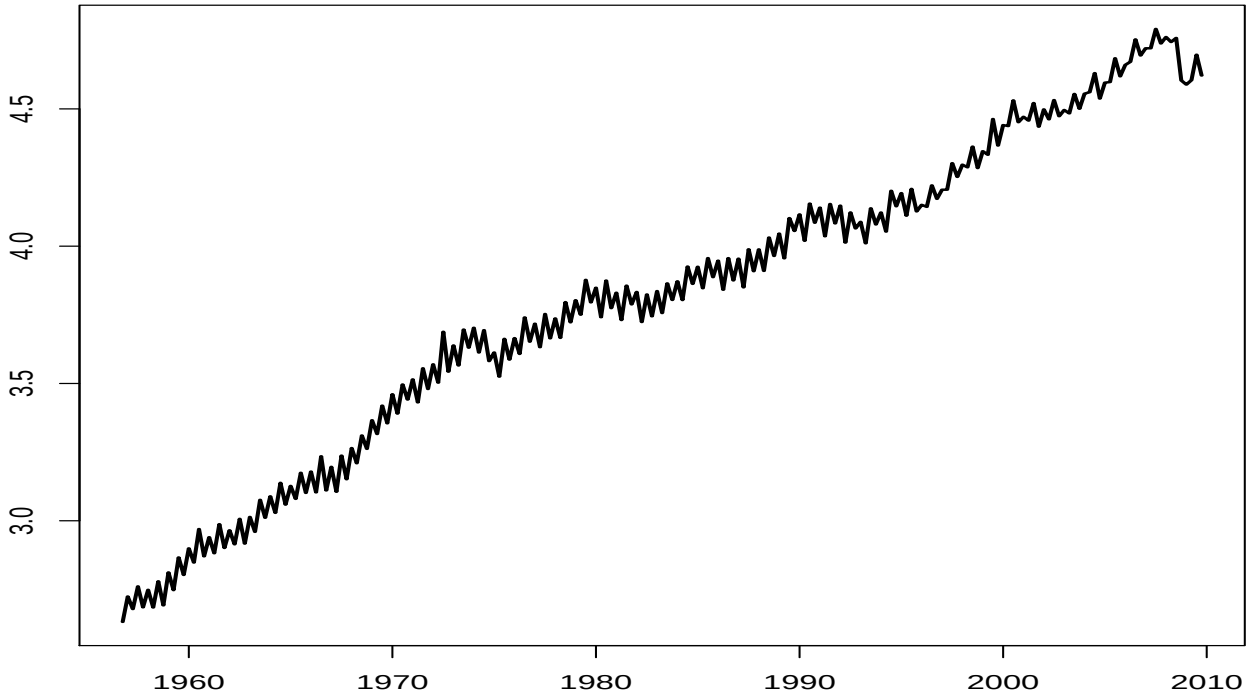


The Wolf&Wölfer sunspot numbers, annual data 1770–1869 (from Brockwell&Davis).

Examples of time series: Accidental deaths

Monthly accidental deaths in the USA, 1973–1978 (from Brockwell&Davis, p.7).

Examples of time series: Industrial production



Austrian industrial production, quarterly observations 1958–2009.

Stochastic processes

Definition

A (discrete-time) *stochastic process* (X_t) is a sequence of random variables that is defined on a common probability space.

Remark. The index set $(t \in \mathbb{I})$ is 'time' and is $\mathbb{N}$ or $\mathbb{Z}$. The random variables may be real-valued (univariate time series) or real-vector valued (multivariate time series). This time-series process is sometimes also called 'time series'.

Examples of stochastic processes: random numbers

1. Rolling a die yields a stochastic process with independent, discrete-uniform observations on the codomain $\{1, \ldots, 6\}$.
2. Tossing a coin yields a stochastic process with independent, discrete-uniform observations on the codomain $\{\mathit{heads}, \mathit{tails}\}$.
3. This can be generalized to a random-number generator (any distribution, typically continuous) on a computer. However, strictly speaking, the outcome is not random ('pseudo random numbers').

Remark. These are very simple stochastic processes that do not have an interesting structure.

The idea of a random walk

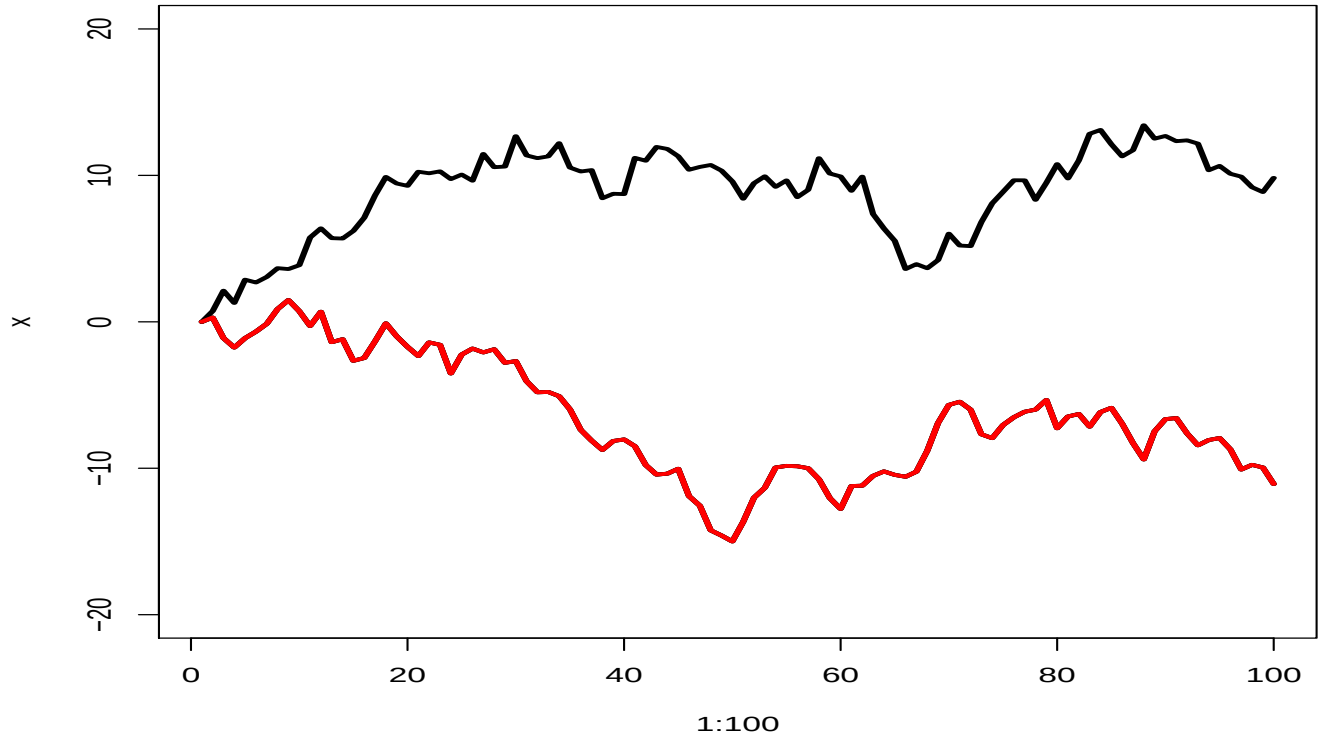
Start in time $t = 0$ at a fixed value X_0 and define successive values for $t > 0$ by

$$X_t = X_{t-1} + u_t,$$

where u_t is purely random (for example, uniform on $\{-1, 1\}$ for a drunk man's steps or normal for the usual random walk in economics). Now, the process has dependence structure over time.

This important process is called the *random walk* (RW).

Two realizations of the random walk

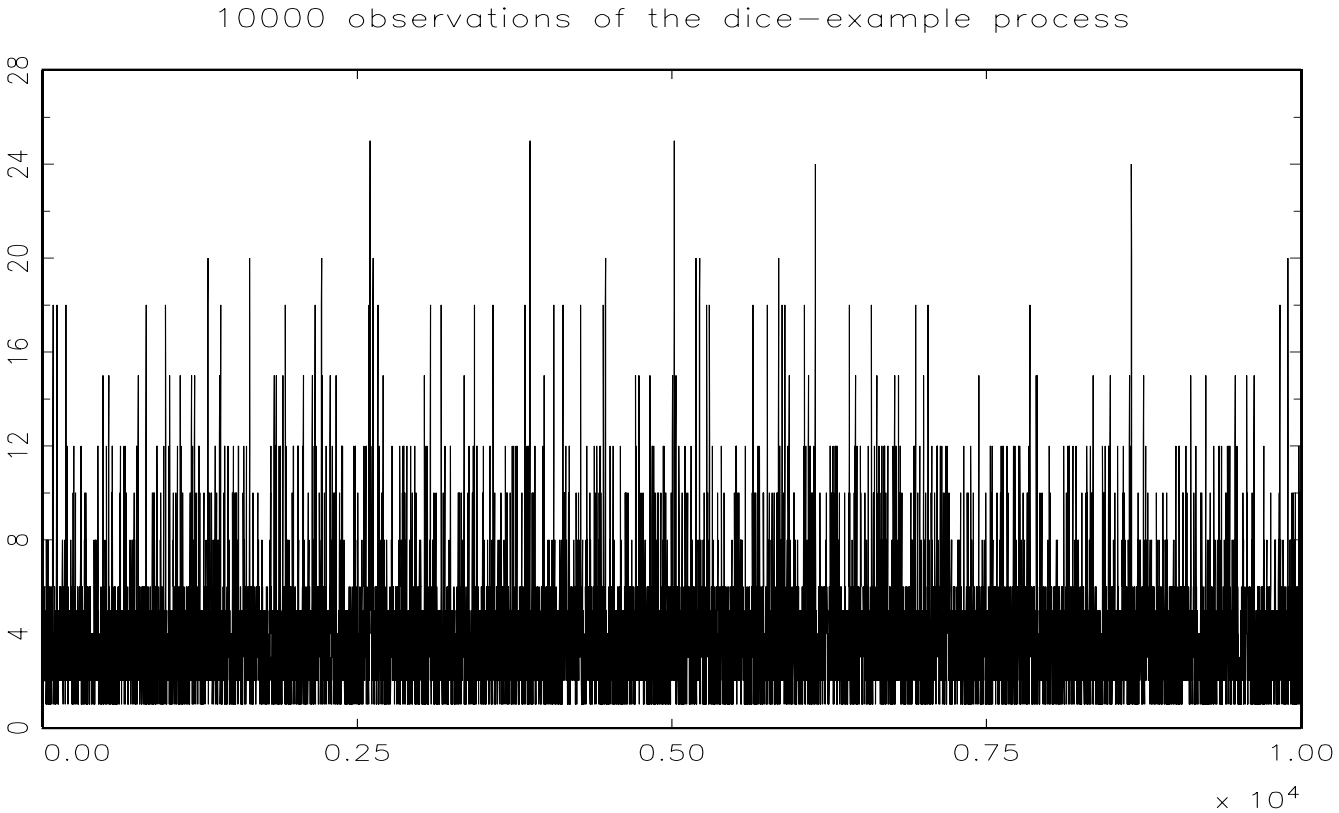


Random walk with standard normal increments, started in $X_0 = 0$.

Examples of stochastic processes: rolling a die with bonus

Suppose that you roll a die and every time that the face with the number six is up, the next roll counts double score; if you get another six, the next roll counts triple etc. Then, there is dependence over time, and the process is not even time-reversible. Nonetheless, the first 100 observations will look approximately like the next 100.

Time plot of rolling with bonus



10,000 rolls of a die with bonus for face 6.

Stationarity

Definition

A time-series process (X_t) is called *covariance stationary* or *weakly stationary* iff

$$\begin{aligned} \mathrm{E}X_t &= \mu \quad \forall t, \\ \mathrm{var}X_t &= \sigma^2 \quad \forall t, \\ \mathrm{cov}(X_t, X_{t-h}) &= \mathrm{cov}(X_s, X_{s-h}) \quad \forall t, s, h. \end{aligned}$$

The process is called *strictly stationary* if the joint distribution of any finite sequence of observations does not change over time. In the following, only weak stationarity will be needed.

Covariance stationarity and strict stationarity

Strict stationarity implies covariance stationarity if variance is finite. Covariance stationarity implies strict stationarity if all distributions are normal.

Examples. Random draws from a heavy-tailed distribution with infinite variance are strictly but not weakly stationary. Drawing from different distributions with identical variance and mean yields a weakly stationary but not strictly stationary process.

Stationary or not?

- Rolling a die (even with bonus) defines a strictly and covariance stationary process;
- A random walk is not stationary: variance increases over time;
- Sunspots may be stationary;
- The U.S. population may be non-stationary.

White noise

Definition

A stochastic process (ε_t) is called a *white noise* iff

1. $\mathrm{E}\varepsilon_t = 0 \quad \forall t;$
2. $\mathrm{var}\varepsilon_t = \sigma^2 \quad \forall t;$
3. $\mathrm{cov}(\varepsilon_t, \varepsilon_s) = 0 \quad \forall s \neq t.$

Remark. A white noise may have time-changing distributions, nonlinear or higher-order dependence over time, and even its mean may be nontrivially predictable. A white noise need not be *iid* (independent identically distributed).

Examples for white noise

1. Random draws are *iid*. If they have a finite variance and a zero mean, they are white noise;
2. Draws from different distributions with identical expectation 0 and identical variance are white noise but not *iid*.

Note. Random draws from a distribution with infinite variance (Pareto or Cauchy) are iid but not white noise.

The autocorrelation function

Weakly stationary processes are characterized by their autocovariance function (ACVF)

$$C(h) = \mathrm{cov}(X_t, X_{t+h}),$$

with $C(0) = \mathrm{var} X_t$ or (more commonly) by their autocorrelation function (ACF)

$$\rho(h) = \frac{\mathrm{E}\{(X_t - \mu)(X_{t+h} - \mu)\}}{\mathrm{E}\{(X_t - \mu)^2\}} = \frac{C(h)}{C(0)},$$

with $\rho(0) = 1$.

Note that $\mathrm{var} X_t = \mathrm{var} X_{t+h}$ by stationarity, so the function really evaluates correlations. The sample ACF is called a 'correlogram'.

Properties of the ACF

- $\rho(h) = \rho(-h)$. Therefore, the ACF is often considered just for $h \geq 0$;
- $|\rho(h)| \leq 1 \quad \forall h$;
- $\mathrm{corr}(X_t, X_{t+1}, \ldots, X_{t+h})$ must be a correlation matrix and nonnegative definite: $\rho(h)$ is a non-negative definite function.

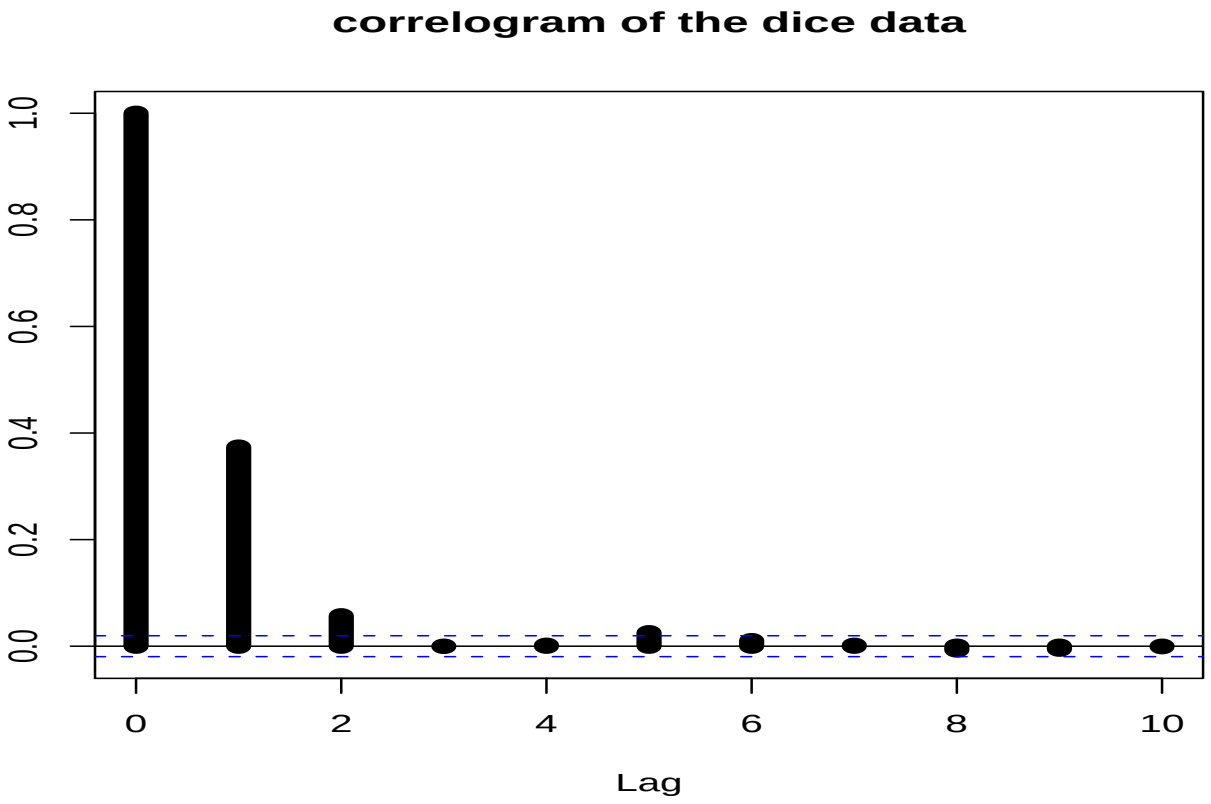
Remark. It can be shown that for any given non-negative definite function there is a stationary process with the given function as its ACF.

Examples for the ACF

White noise has a trivial ACF: $\rho(0) = 1$, $\rho(h) = 0, h > 0$.

Rolling a die with bonus has a non-trivial ACF. There is a stationary process that would match the ACF without revealing the generating law. That process and the bonus-roll process are then seen as equivalent.

Correlogram of rolling with bonus



Trend and seasonality

Many time series show regular, systematic deviations from stationarity: trend and seasonal variation. It is often possible to transform such variables to stationarity.

Other deviations (breaks, outliers etc.) are more difficult to handle.

Trend-removing transformations

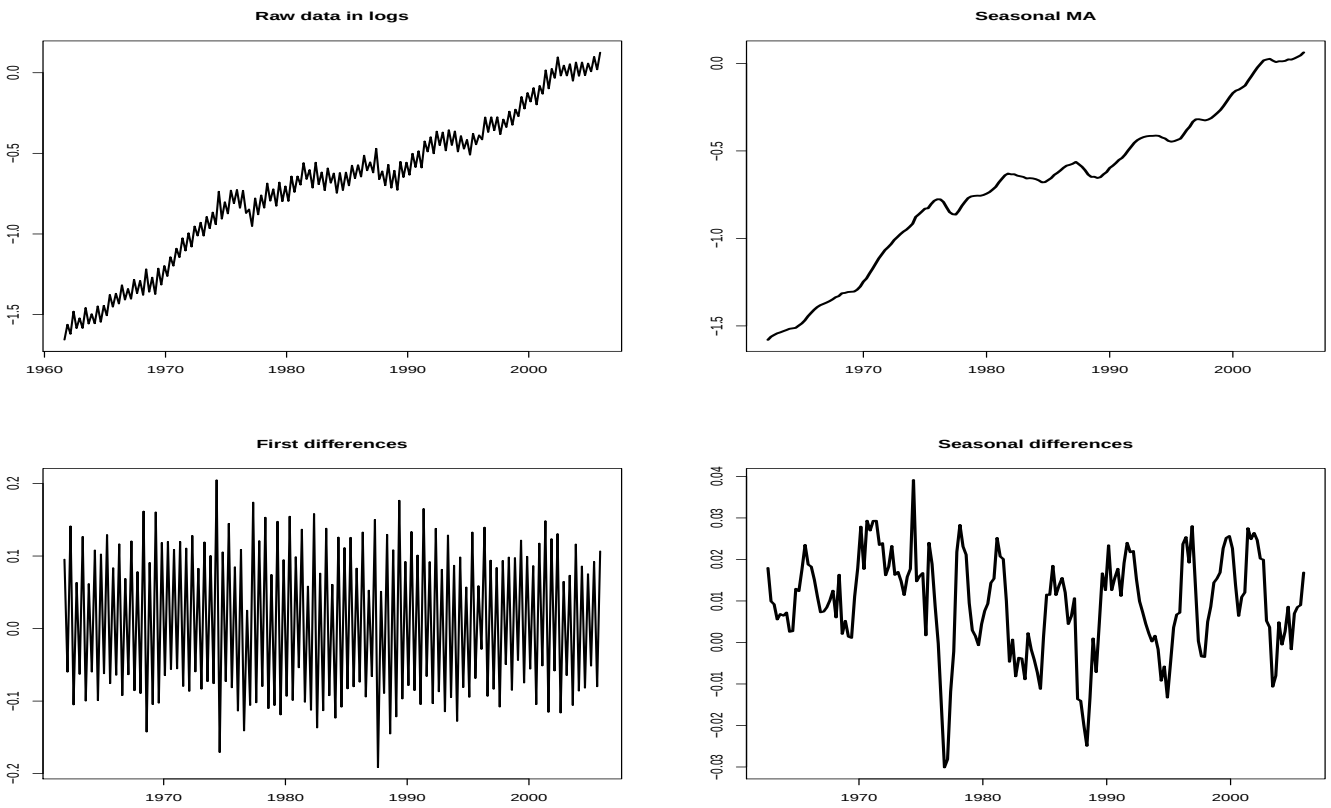
1. Regression on simple functions of time (linear, quadratic) such that residuals are approximately stationary. Often, after logarithmic preliminary transforms (exponential trend): assumes *trend stationarity*;
2. First-order differences $X_t - X_{t-1} = \Delta X_t$. Often, after logarithmic transformations (logarithmic growth rates): assumes *difference stationarity*.

Remark. Trend regressions assume a time-constant trend structure that may not be plausible.

Seasonality-removing transformations

1. Regression on seasonal dummy constants and proceeding with the residuals (assumes time-constant seasonal pattern);
2. Seasonal adjustment (Census X–11, X–12, TRAMO-SEATS): nonlinear irreversible transformations;
3. Seasonal differencing $X_t - X_{t-s} = \Delta_s X_t$: assumes time-changing seasonal patterns and removes trend and seasonality jointly;
4. Seasonal moving averages $X_t + X_{t-1} + \ldots + X_{t-s+1}$ maybe multiplied with $1/s$: seasonal differences without trend removal.

Example for transformations: Austrian industrial production



Moving-average processes

Definition

Assume (ε_t) is white noise. The process (X_t) given by

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

is a *moving-average process of order* q or, in short, an MA(q) process.

The MA(1) process $X_t = \varepsilon_t + \theta \varepsilon_{t-1}$ is the simplest generalization of the white noise that allows for dependence over time.

Why MA models?

Wold's Theorem states that every covariance-stationary process (X_t) can be approximated to an arbitrary degree of precision by an MA process (of arbitrarily large order q) plus a deterministic part (constant, maybe sine waves of time):

$$X_t = \delta_t + \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}$$

The white noise in the MA representation is constructively defined as the forecast error in linearly predicting X_t from its past and is also named *innovations*.

Basic properties of MA processes

1. An MA process defined on $t \in \mathbb{Z}$ is stationary. If defined on $t \in \mathbb{N}$, it is stationary except for the first q observations;
2. An MA(q) process is linearly q–dependent, $\rho(h) = 0$ for $h > q$;
3. Values of the parameters θ_j are not restricted to any area of $\mathbb{R}$. However, then it is not uniquely defined. In order to get the MA process of the Wold Theorem, it is advisable to constrain the values.

The identification problem in MA models

The MA(1) processes $X_t = \varepsilon_t + \theta\varepsilon_{t-1}$ and $X_t^* = \varepsilon_t + \theta^*\varepsilon_{t-1}$ with $\theta^* = \theta^{-1}$ have the same ACF: $\rho(1) = \rho^*(1) = \theta/(1+\theta^2)$. Observing data does not permit deciding whether θ or θ^* has generated the data: lack of identifiability.

Larger q even aggravates the identification problem.

Identification of general MA(q) models

For the MA(q) model

$$X_t = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j},$$

o.c.s. that the *characteristic polynomial*

$$\Theta(z) = 1 + \sum_{j=1}^{q} \theta_j z^j$$

determines all equivalent structures. If $\zeta \in \mathbb{C}$ is any root (zero) of $\Theta(z)$, it can be replaced by ζ^{-1} without changing the ACF. There are up to 2^q equivalent MA models (note potential multiplicity, complex conjugates).

Unique definition of MA(q) models

If the characteristic polynomial $\Theta(z) = 1 + \sum_{j=1}^{q} \theta_j z^j$ has only roots ζ with $|\zeta| \geq 1$, the MA(q) model is identified and uniquely defined.

Why not small roots? This definition corresponds to the MA representation in Wold's Theorem. Large roots mean small coefficients. If all roots fulfil $|\zeta| > 1$, the MA model is *invertible* and there is an autoregressive representation

$$X_t = \sum_{j=1}^{\infty} \phi_j X_{t-j} + \varepsilon_t,$$

which may be useful for prediction.

The ACF of an MA process

It is easy to evaluate the ACF of a given MA process and to derive the formula

$$\rho(h) = \frac{\sum_{i=0}^{q-h} \theta_i \theta_{i+h}}{\sum_{i=0}^{q} \theta_i^2}, \quad h \leq q,$$

using formally $\theta_0 = 1$. For $h > q$, $\rho(h) = 0$. For small q, the intensity of autocorrelation is constrained to comparatively small values: MA is not a good model for observed strong autocorrelation.

Infinite-order MA processes

Wold's Theorem expresses the non-deterministic part of a stationary process essentially as an infinite-order MA process

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

which converges in mean square and hence in probability, as $\sum_{j=0}^{\infty} \theta_j^2 < \infty$. O.c.s. that convergence of

$$\sum_{j=0}^{\infty} a_j X_{t-j}$$

for an arbitrary stationary process X_t will hold if $\sum_{j=0}^{\infty} |a_j| < \infty$.

For an infinite sum of non-white noise, the condition is slightly stronger.

The first-order autoregressive model

Definition

The process (X_t) constructed from the equation

$$X_t = \phi X_{t-1} + \varepsilon_t,$$

with (ε_t) a white noise (ε_t uncorrelated with X_{t-1}), and $\phi \in \mathbb{R}$, is called a *first-order autoregressive process* or AR(1). Depending on the value of ϕ, it may be defined for $t \in \mathbb{Z}$ or for $t \in \mathbb{N}$.

The case $|\phi| < 1$

Repeated substitution into the AR(1) equation yields

$$X_t = \phi^t X_0 + \sum_{j=0}^{t-1} \phi^j \varepsilon_{t-j}$$

for $t \in \mathbb{N}$. For $t \in \mathbb{Z}$, substitution can be continued to yield

$$X_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j},$$

which converges as $\sum_{j=0}^{\infty} \phi^{2j} = 1/(1-\phi^2)$ converges.

For $\mathbb{Z}$, the resulting process is easily shown to be stationary. For $\mathbb{N}$, it is not stationary but becomes stationary for $t \to \infty$: 'asymptotically stationary' or 'stable'.

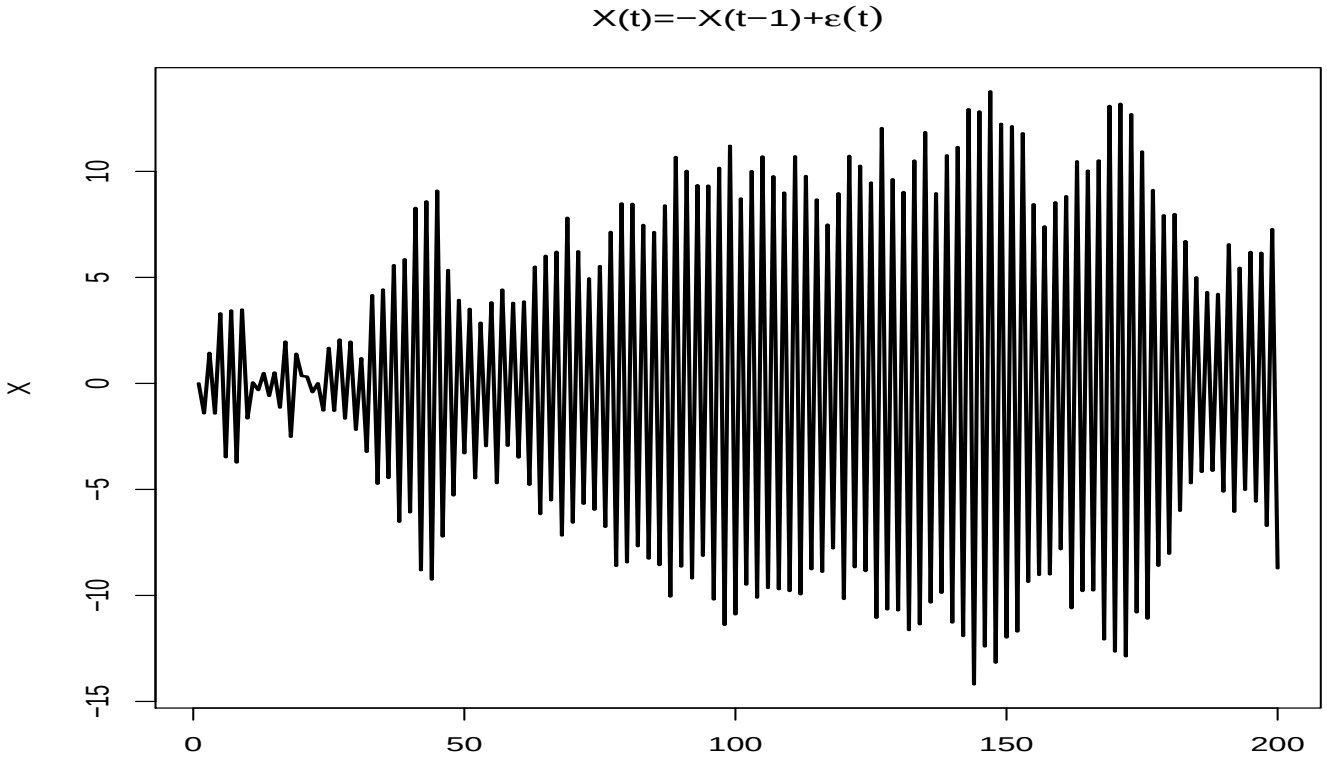
The case $|\phi| = 1$

$\phi = 1$ defines $X_t = X_{t-1} + \varepsilon_t$, the random walk, which is non-stationary. $\phi = -1$ implies $X_t = -X_{t-1} + \varepsilon_t$, a very similar process, jumping between positive and negative values, sometimes called the 'random jump'. Because of

$$X_t = X_0 + \sum_{j=0}^{t-1} (-1)^j \varepsilon_{t-j},$$

$\mathrm{var} X_t = t\sigma_\varepsilon^2$, and the process cannot be stationary.

Time plot of a random jump



The case $|\phi| > 1$

If $|\phi| > 1$, repeated substitution yields

$$X_t = \phi^t X_0 + \sum_{j=0}^{t-1} \phi^j \varepsilon_{t-j}$$

for $t \in \mathbb{N}$. The first term increases exponentially as $t \to \infty$, and the process cannot be stationary. These processes are called *explosive*.

By reverse substitution, a stationary process may be defined that lets the past depend on the future. Such 'non-causal' processes violate intuition and will be excluded.

The variance of a stationary AR(1) process

If $X_t = \phi X_{t-1} + \varepsilon_t$ is stationary, clearly $\mathrm{E}X_t = 0$, $\mathrm{var}(X_t) = \mathrm{var}(X_{t-1})$, and hence

$$\begin{aligned} \mathrm{var}X_t &= \mathrm{var}(\phi X_{t-1} + \varepsilon_t) \\ &= \phi^2 \mathrm{var}(X_t) + \sigma_\varepsilon^2, \end{aligned}$$

which yields

$$\mathrm{var}X_t = C(0) = \frac{\sigma_\varepsilon^2}{1 - \phi^2}.$$

The ACF of a stationary AR(1) process

If (X_t) is stationary,

$$\begin{aligned} C(1) = \mathrm{E}(X_t X_{t-1}) &= \mathrm{E}\{X_{t-1}(\phi X_{t-1} + \varepsilon_t)\} \\ &= \phi C(0). \end{aligned}$$

Similarly, we obtain $C(h) = \phi^h C(0)$ for all larger h, which also implies $\rho(h) = \phi^h$ for all h. The ACF 'decays' geometrically to 0 as $h \to \infty$.

The autoregressive model of order p

Definition

The process (X_t) constructed from the equation

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \varepsilon_t,$$

with (ε_t) a white noise (ε_t uncorrelated with X_{t-1}), and $\phi \in \mathbb{R}$, is called an *autoregressive process of order* p or AR(p). Depending on the values of its coefficients, it may be defined for $\mathbb{I} = \mathbb{Z}$ or for $\mathbb{I} = \mathbb{N}$.

Remark. For higher lag orders p, direct stability conditions on coefficients become insurmountably complex. It is more convenient to apply the following theorem on the *characteristic polynomial*.

Stability of AR(p) processes

Theorem

The AR(p) process

$$X_t = \mu + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t$$

has a covariance-stationary causal solution (is stable) if and only if all roots ζ_j of the characteristic polynomial

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p$$

fulfil the condition $|\zeta_j| > 1$, i.e. they are all situated 'outside the unit circle'.

The ACF of a stationary AR(p) process

If the AR(p) model is stationary, it has an infinite-order MA representation

$$X_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j},$$

with geometrically declining weights θ_j. The rate of convergence to 0 is determined by the inverse of the root of the polynomial $\Phi(z)$ that is closest to the unit circle.

For this reason, the ACF will show a geometric decay as $h \to \infty$, following an arbitrary pattern of p starting values. It is quite difficult to read the true order of an AR(p) process off the correlogram.

The partial autocorrelation function

There are two equivalent definitions for the *partial autocorrelation function* (PACF). In the first, AR(p) models for $p = 1, 2, \ldots$ are fitted to data:

$$
\begin{aligned}
X_t &= \phi_1^{(1)} X_{t-1} + u_t^{(1)}, \\
X_t &= \phi_1^{(2)} X_{t-1} + \phi_2^{(2)} X_{t-2} + u_t^{(2)}, \\
X_t &= \phi_1^{(3)} X_{t-1} + \phi_2^{(3)} X_{t-2} + \phi_3^{(3)} X_{t-3} + u_t^{(3)}, \\
&\ldots \\
X_t &= \phi_1^{(p)} X_{t-1} + \phi_2^{(p)} X_{t-2} + \ldots + \phi_p^{(p)} X_{t-p} + u_t^{(p)}.
\end{aligned}
$$

The PACF is defined as the limits of the last coefficient estimates $\phi_1^{(1)}, \phi_2^{(2)}, \ldots$

The cutoff property of the PACF for AR(p) models

If the generating process is AR(p), then:

1. For $j < p$, disturbance terms $u_t^{(j)}$ are not white noise. Coefficient estimates typically converge to non-zero limits;
2. For $j = p$, $u_t^{(p)}$ is white noise for the true coefficients, which are limits of the consistent estimates. The last coefficient estimate $\hat{\phi}_p^{(p)}$ converges to a non-zero ϕ_p.
3. For $j > p$, the models are over-fitted. The first p coefficient estimates converge to the true ones, the others and particularly the last one converge to 0.

The PACF will be non-zero for $j \leq p$ and zero for $j > p$. It cuts off at p and thus indicates the true order.

PACF and ACF for AR and MA

O.c.s. that the PACF for an MA process decays to zero geometrically. This is plausible, as no AR process really fits and invertible MA models have infinite-order AR representations with geometrically declining weights. This result closes the simple 'duality':

1. MA(q) processes have an ACF that cuts off at q and a geometrically declining PACF;
2. AR(p) processes have a PACF that cuts off at p and a geometrically declining ACF;
3. There are also processes whose ACF and PACF both decay smoothly: the ARMA processes (next topic).

The definition of ARMA models

Definition

The process (X_t) defined by the ARMA model

$$X_t = \mu + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

is called an ARMA(p, q) process, provided it is stable (asymptotically stationary) and uniquely defined.

Remark. In line with many authors in time series, unstable ARMA models do not define ARMA processes. A random walk is not an ARMA process.

Conditions for unique and stable ARMA

Theorem

An ARMA process is uniquely defined and stable iff

1. *the characteristic AR polynomial* $\Phi(z)$ *has roots* ζ *for* $|\zeta| > 1$ *only;*
2. *the characteristic MA polynomial* $\Theta(z)$ *has roots* ζ *for* $|\zeta| \geq 1$ *only;*
3. *the polynomials* $\Phi(z)$ *and* $\Theta(z)$ *have no common roots.*

Remarks. Condition (1) establishes stability, which is unaffected by the MA part. Condition (2) implies uniqueness of the MA part and invertibility if there are no roots with $|\zeta| = 1$. Condition (3) implies uniqueness of the entire structure.

Cancelling roots in ARMA models

Suppose (X_t) is white noise, i.e.

$$X_t = \varepsilon_t \therefore X_{t-1} = \varepsilon_{t-1},$$

and hence for any ϕ

$$X_t - \phi X_{t-1} = \varepsilon_t - \phi \varepsilon_{t-1},$$

apparently an ARMA(1,1) model with redundant parameter ϕ. Note that $\Phi(z) = \Theta(z) = 1 - \phi z$. Condition (3) excludes such cases.

Determining the lag orders p and q

If (X_t) is ARMA(p, q), ACF and PACF show a geometric decay that starts more or less from q and p. It is almost impossible to recognize lag orders from the correlogram reliably.

Generalization of ACF and PACF that show p and q clearly (corner method, extended ACF) are rarely used. It is common to

1. estimate various ARMA(p, q) models for a whole range $0 \leq p \leq P$ and $0 \leq q \leq Q$;
2. compare all $(P + 1)(Q + 1)$ models by information criteria (IC) and pick the model with smallest IC;
3. subject the selected model to specification tests.

If the model turns out to be specified incorrectly, most authors would try out another model with larger p and/or q.

Information criteria

If an ARMA(p, q) model is estimated, label the variance estimate of its errors ε_t as $\hat{\sigma}^2(p, q)$. Then,

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

define the most popular information criteria. Both were introduced by Akaike, BIC was modified by Schwarz.

Information criteria and hypothesis tests

There is no contradiction between IC and hypothesis tests.

- Basically, IC add a penalty term to the negative log-likelihood. The log-likelihood tends to increase if the model becomes more sophisticated (larger p and q). The penalty term prevents the models from becoming too 'profligate'.
- An LR test does the same. It compares the log-likelihood difference between two models to critical values.
- IC automatically define the significance level of a comparable hypothesis test. For BIC, this level converges to 0 as $T \to \infty$.
- The IC approach can compare arbitrarily many models that may be non-nested. Hypothesis testing can compare two nested models only.

Some common misunderstandings in time-series applications

1. Distribution of coefficient tests: coefficient estimates divided by standard errors are not t–distributed under their null but still asymptotically standard normal N(0,1).
2. Durbin-Watson test: this is a test for autocorrelation in the errors of a static regression. In a dynamic model, such as ARMA or even AR, it is not meaningful.
3. Jarque-Bera test: a convenient check for the normality of errors. However, errors need not be normally distributed in a correct model. Increasing the lag order usually does not yield normal errors.

Whiteness checks

A well-specified ARMA model has white-noise errors. Several tests are appropriate:

1. Correlogram plots of the residuals should show no too significant values.
2. The portmanteau Q by Ljung and Box summarizes the correlogram as

$$Q = T(T+2)\sum_{j=1}^{J}\frac{\hat{\rho}_j^2}{T-j} \quad .$$

   Under the null of a white noise, it is asymptotically distributed as $\chi^2(J-p-q)$. This test has low power.
3. The usual LM tests for autocorrelation in errors can be used. They are similar to the Q test.

It is not advisable to determine the lag orders p and q based on these whiteness checks only.