Introduction to Time Series Analysis

Dr David A. Stephens

Department of Mathematics & Statistics
McGill University
Room 1225, Burnside Hall

david.stephens@mcgill.ca
www.math.mcgill.ca/dstephens/

Time Series: Prerequisites

- Basic distribution theory
- Expectations and other moments
- Basics of statistical theory (ie estimation, testing, large sample theory)
- Linear algebra (matrix manipulations and decomposition)
- Basic knowledge of complex numbers
- Computing an advantage (knowledge of R).

Time Series: Objectives

The basic objective is to construct models for observed data

$$x_1, x_2, \ldots, x_t, \ldots$$

with t indexing time, with the goals of

- explaining the underlying data generating mechanism
- forecasting future values of the underlying process.

We will mainly focus on discrete time modelling, and denote the underlying stochastic (random) process $\{X_t\}$.

Time Series: Objectives

The stochastic process can be (partly) characterized via its

- finite dimensional distributions, that is the joint distributions of

  $$(X_{t_1}, \ldots, X_{t_n})$$

  for each index set $\{t_1, \ldots, t_n\}$
- moment properties (expectation, variance, covariance)

Typically $\{X_t\}$ will not be independent, and inference about the parameters of the joint distribution will be of central interest.

Time Series: Forecasting

Given observed values $X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T$, we aim to predict or forecast future values of the process

$$X_{T+1}, \ldots, X_{T+K}$$

say.

For example, could use a simple linear predictor

$$\hat{X}_{T+1} = \sum_{t=1}^{T} a_t x_t$$

that is, a linear combination of past values of the process. The objective then is to choose the optimal values for the a_t.

Chapter 1
Introduction

1.1 Introduction

A time series is a sequence of random variables

$$\{X_t\}_{t=1,\ldots,T} = (X_1, \ldots, X_T) = X_{1:T}$$

or a realization

$$\{x_t\}_{t=1,\ldots,T} = (x_1, \ldots, x_T) = x_{1:T}$$

collected over time index t representing some unit time period (hours, days, months etc.).

1.2 Simple Stochastic Models

A time series model is a probability model specifying

- the joint distribution
- the second-order moment structure (mean, variance, covariance)

of $\{X_t\}$.

The joint distribution specifies the moment structure.

1.2 Simple Stochastic Models

Recall that for any two random variables, the covariance is defined as

$$\text{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

where $\mu_i = \mathbb{E}[X_i]$ and $\mu_j = \mathbb{E}[X_j]$, and the correlation is

$$\text{Corr}[X_i, X_j] = \frac{\text{Cov}[X_i, X_j]}{\sqrt{\text{Var}[X_i]\text{Var}[X_j]}}$$

1.2 Simple Stochastic Models

For a d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$, we form the covariance matrix, Σ, so that the (i,j)th element of Σ is

$$[\Sigma]_{i,j} = \mathrm{Cov}[X_i, X_j].$$

This matrix is symmetric, and non-negative definite. We may write

$$\Sigma = \mathbf{VRV}^\top$$

where

$$\mathbf{V} = \mathrm{diag}(\sqrt{\mathrm{Var}[X_1]}, \ldots, \sqrt{\mathrm{Var}[X_d]})$$

and $\mathbf{R}$ is the matrix of correlations.

1.2 Simple Stochastic Models

Special Case: Independence

$$F_{X_{1:T}}(x_{1:T}) = \prod_{t=1}^{T} F_X(x_t) = \prod_{t=1}^{T} \Pr[X_t \leqslant x_t]$$

or equivalently, for each t,

$$\mathbb{P}[X_t \leqslant x_t | X_{1:(t-1)} = x_{1:(t-1)}] = \Pr[X_t \leqslant x_t]$$

where the past values of X contain no information about future X.

Note: If two variables are independent, then they are also uncorrelated; however, in general uncorrelatedness does not imply independence.

1.2 Simple Stochastic Models

For the specification via moments, interest focusses on

$$\begin{aligned} \text{Expectation} \quad & \mu_t = \mathbb{E}[X_t] \\ \text{Variance} \quad & \sigma^2_{X_t} = \text{Var}[X_t] \\ \text{Covariance} \quad & \gamma_X(t,s) = \mathbb{E}[(X_t - \mu_t)(X_s - \mu_s)] \end{aligned}$$

for any indices t, s. As the X_t usually represent repeated measurements of the same phenomenon over time, the term autocovariance is used.

1.2 Simple Stochastic Models

Example: IID Process

Let $0 < p < 1$, and suppose

$$\mathbb{P}[X_t = 1] = p \qquad \mathbb{P}[X_t = -1] = 1 - p$$

with the X_t mutually independent. Then $\{X_t\}$ is an IID (independent, identically distributed) process.

1.2 Simple Stochastic Models

Example: IID Process

We have that

$$\mu_t = p + (1-p)(-1) = 2p - 1$$

$$\sigma_t^2 = [p(1)^2 + (1-p)(-1)^2] - (2p-1)^2 = 1 - (1-2p)^2$$

$$\gamma(t,s) = 0 \qquad t \neq s$$

and these quantities do not depend non t or s.

1.2 Simple Stochastic Models

Example: Random Walk

Suppose $\{X_t\}$ is an IID process, and let

$$S_t = \sum_{i=1}^{t} X_i = S_{t-1} + X_t$$

Then $\{S_t\}$ is a random walk. Also, setting $S_0 = 0$, we have

$$X_t = S_t - S_{t-1}$$

that is, $\{X_t\}$ can be obtained from $\{S_t\}$ by differencing.

1.2 Simple Stochastic Models

Example: Random walk

Suppose $\{X_t\}$ is an IID process with

$$\mathbb{P}[X_t = 1] = \mathbb{P}[X_t = -1] = \frac{1}{2}$$

and zero otherwise, and let

$$S_t = \sum_{i=1}^{t} X_i.$$

Then both $\{X_t\}$ and $\{S_t\}$ are zero-mean processes.

If $p \neq 1/2$, then $\mathbb{E}[X_t], \mathbb{E}[S_t]$ are non-zero; if $p < 1/2$, then $\mathbb{E}[S_t] < 0$ and we have a downward drift.

Example: Random walk

Note that for any (time-homogeneous) random walk

$$\mathrm{Var}[S_t] = \mathrm{Var}\left[\sum_{i=1}^{t} X_i\right] = \sum_{i=1}^{t} \mathrm{Var}\left[X_i\right] = t\mathrm{Var}\left[X_1\right]$$

by the independence of the X_t, so the variance grows linearly with t.

Thus even if $\mathbb{E}[S_t] = 0$, the probability distribution of S_t has increasing variability around zero, and will (almost surely) diverge to an infinite value.

1.3 Gaussian processes

Suppose, for all finite collections $X_1, \ldots, X_n$, $n \geqslant 1$, the joint distribution is multivariate Gaussian (Normal) with

$$\mathbb{E}[X_t] = \mu_X$$

and covariance defined for X_t, X_s as

$$\mathrm{Cov}[X_t, X_s] = \gamma_X(|t - s|) \qquad t, s \in \{1, \ldots, n\}.$$

for some function of a single argument γ_X. We have that

$$\mathrm{Var}[X_t] = \gamma_X(0).$$

1.3 Gaussian processes

Note that the autocovariance only depends on the value of

$$|t - s|$$

This version imposes extra conditions, which corresponds to a structured covariance matrix which has fewer than $n(n+1)/2$ different elements.

1.3 Gaussian processes

Let $\mathbf{\Gamma}_X(n)$ denote the $(n \times n)$ matrix with

$$[\mathbf{\Gamma}_X(n)]_{t,s} = \gamma_X(|t-s|).$$

Then $\mathbf{\Gamma}_X(n)$ is a symmetric, positive definite matrix so that

$$\mathbf{x}^\top \mathbf{\Gamma}_X(n)\mathbf{x} > 0 \qquad \mathbf{x} \in \mathbb{R}^n.$$

1.3 Gaussian processes

The matrix $\mathbf{\Gamma}_X(n)$ has Toeplitz structure, constant among diagonals,

$$\mathbf{\Gamma}_X(n) = \begin{bmatrix} \gamma_X(0) & \gamma_X(1) & \cdots & \cdots & \gamma_X(n-1) \\ \gamma_X(1) & \gamma_X(0) & \gamma_X(1) & \cdots & \vdots \\ \gamma_X(2) & \gamma_X(1) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \gamma_X(1) \\ \gamma_X(n-1) & \cdots & \cdots & \gamma_X(1) & \gamma_X(0) \end{bmatrix}$$

We write the vector

$$\mathbf{X}_n \equiv X_{1:n} \sim \mathcal{N}_n(\mu_X \mathbf{1}_n, \mathbf{\Gamma}_X(n)).$$

where $\mathbf{1} = (1, 1, \ldots, 1)^\top$ is an $n \times 1$ vector.

1.3 Gaussian processes

We have the decomposition

$$\mathbf{X}_n = \mu_X \mathbf{1}_n + \mathbf{Z}_n \qquad (n \times 1)$$

where

$$\mathbf{Z}_n \sim \mathcal{N}_n(\mathbf{0}_n, \mathbf{\Gamma}_X(n)).$$

1.3 Gaussian processes

In general, if

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

with $\mathbf{X}_1$ and $\mathbf{X}_2$ being $k_1 \times 1$ and $k_2 \times$ respectively, where

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

then

$$\mathbf{X}_1 \sim \mathcal{N}_{k_1}\left(\boldsymbol{\mu}_1, \Sigma_{11}\right)$$

$$\mathbf{X}_2 | \mathbf{X}_1 = \mathbf{x}_1 \sim \mathcal{N}_{k_2}\left(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right).$$

1.3 Gaussian processes

Specifically, if we consider the one-step-ahead (conditional) probability distribution

$$X_{n+1} | X_{1:n} = x_{1:n}$$

it follows by standard properties that

$$X_{n+1} | X_{1:n} = x_{1:n} \sim \mathcal{N}\left(\mu_{(n+1)|n}, \sigma^2_{(n+1)|n}\right)$$

1.3 Gaussian processes

Write $\gamma_k = \gamma_X(k)$ for each $k \geqslant 0$. The covariance matrix of $X_{1:(n+1)}$ is

$$\boldsymbol{\Gamma}_X(n+1) = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-2} & \gamma_{n-1} & \gamma_n \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-3} & \gamma_{n-2} & \gamma_{n-1} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{n-4} & \gamma_{n-3} & \gamma_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \gamma_{n-2} & \gamma_{n-3} & \gamma_{n-4} & \cdots & \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_{n-1} & \gamma_{n-2} & \gamma_{n-3} & \cdots & \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_n & \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{\Gamma}_X(n) & \boldsymbol{\gamma}_n^{\mathrm{R}} \\ \{\boldsymbol{\gamma}_n^{\mathrm{R}}\}^\top & \gamma_0 \end{bmatrix}$$

where

$$\boldsymbol{\gamma}_n^{\mathrm{R}} = [\gamma_n, \gamma_{n-1}, \gamma_{n-1}, \ldots, \gamma_2, \gamma_1]^\top$$

1.3 Gaussian processes

Applying the previous formula, we have

$$\mu_{(n+1)|n} = \mu_X + \{\boldsymbol{\gamma}_n^{\mathrm{R}}\}^\top \{\boldsymbol{\Gamma}_X(n)\}^{-1}(\mathbf{x}_{1:n} - \mu_X \mathbf{1}_n),$$

and

$$\sigma^2_{(n+1)|n} = \gamma_0 - \{\boldsymbol{\gamma}_n^{\mathrm{R}}\}^\top \{\boldsymbol{\Gamma}_X(n)\}^{-1} \boldsymbol{\gamma}_n^{\mathrm{R}}.$$

Note therefore that

$$\sigma^2_{(n+1)|n} < \gamma_0$$

as $\boldsymbol{\Gamma}_X(n)$ is positive definite.

1.4 Stationarity

Suppose $\{X_t\}$ is a time series process, where for all t

$$\mathbb{E}[X_t^2] < \infty.$$

Define

$$\text{Mean function :} \quad \mu_X(t) = \mathbb{E}[X_t]$$

$$\text{Covariance function :} \ \gamma_X(t,s) = \mathbb{E}[(X_t - \mu_X(t))(X_s - \mu_X(s))]$$

for integers t, s. Note that

$$\gamma_X(t,s) = \gamma_X(s,t).$$

1.4 Stationarity

The process $\{X_t\}$ is weakly stationary if

(i) $\mu_X(t)$ does not depend on t

(ii) $\gamma_X(t+h,t)$ does not depend on t for each integer h, so that we may write

$$\gamma_X(t,t+h) \equiv \gamma_X(h).$$

Note also that we must have

$$\gamma_X(-h) = \gamma_X(t,t-h) = \gamma_X(t-h,t) = \gamma_X(h).$$

for $h \geqslant 1$.

Note: The quantity h is often referred to as the lag.

1.4 Stationarity

$\{X_t\}$ is strongly stationary if

$$X_{t:(t+n)} = (X_{t+1}, \ldots, X_{t+n})$$

and

$$X_{(t+h+1):(t+h+n)} = (X_{t+h+1}, \ldots, X_{t+h+n})$$

have the same joint distribution for all t, h, n.

In practice, the (multivariate) distributions required to model such strongly stationary processes are limited in number.

1.4 Stationarity

If $\{X_t\}$ is weakly stationary process define

- the autocovariance function (ACVF), $\gamma_X(h)$, by

$$\gamma_X(h) \equiv \gamma_X(h,0) = \gamma_X(t+h,t)$$

- the autocorrelation function (ACF), $\rho_X(h)$, by

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}$$

That is

$$\gamma_X(h) = \text{Cov}[X_{t+h}, X_t] \qquad \rho_X(h) = \text{Corr}[X_{t+h}, X_t].$$

The fundamental concept of stationarity is that the stochastic structure does not change though time

1.4 Stationarity

Example: IID Process

An IID process $\{X_t\} \sim \text{IID}(0, \sigma_X^2)$, is characterized by

$$\begin{aligned} \mathbb{E}[X_t] &= 0 \\ \text{Var}[X_t] = \mathbb{E}[X_t^2] &= \sigma_X^2 < \infty. \end{aligned}$$

with the X_t mutually independent. This process is stationary:

$$\gamma_X(h) = \text{Cov}[X_{t+h}, X_t] = \begin{cases} \sigma_X^2 & h = 0 \\ 0 & h > 0 \end{cases}$$

$$\rho_X(h) = \text{Corr}[X_{t+h}, X_t] = \begin{cases} 1 & h = 0 \\ 0 & h > 0 \end{cases}$$

1.4 Stationarity

Example: White Noise Process

A white noise process $\{X_t\} \sim \text{WN}(0, \sigma_X^2)$ is characterized by

$$\begin{aligned} \mathbb{E}[X_t] &= 0 \\ \text{Var}[X_t] = \mathbb{E}[X_t^2] &= \sigma_X^2 < \infty. \end{aligned}$$

with the X_t uncorrelated. This process is stationary:

$$\gamma_X(h) = \text{Cov}[X_{t+h}, X_t] = \begin{cases} \sigma_X^2 & h = 0 \\ 0 & h > 0 \end{cases}$$

$$\rho_X(h) = \text{Corr}[X_{t+h}, X_t] = \begin{cases} 1 & h = 0 \\ 0 & h > 0 \end{cases}$$

1.4 Stationarity

Example: Random Walk

If $\{X_t\} \sim \text{WN}(0, \sigma_X^2)$, then the process $\{S_t\}$ defined by

$$S_t = \sum_{i=1}^{t} X_i$$

is non-stationary as, by properties of uncorrelated rvs,

$$\mathbb{E}[S_t] = 0$$

$$\text{Var}[S_t] = t\sigma_X^2$$

which does depend on t.

Example: Random Walk

For the ACVF:

$$\begin{aligned}
\gamma_S(t+h,t) &= \mathrm{Cov}[S_{t+h}, S_t] \\
&= \mathrm{Cov}[S_t + X_{t+1} + \cdots + X_{t+h}, S_t] \\
&= \mathrm{Cov}[S_t, S_t] + \sum_{i=1}^{h} \mathrm{Cov}[X_{t+i}, S_t] \\
&= \mathrm{Cov}[S_t, S_t] \\
&= \mathrm{Var}[S_t] \\
&= t\sigma_X^2.
\end{aligned}$$

Example: Moving Average

Let $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and define process $\{X_t\}$ by

$$X_t = Z_t + \theta_1 Z_{t-1} \qquad t \in \mathbb{Z}$$

where θ_1 is a real-valued parameter. As $\mathbb{E}[Z_t] = 0$, we also have $\mathbb{E}[X_t] = 0$, and by uncorrelatedness of the $\{Z_t\}$ components

$$\text{Var}[X_t] = \text{Var}[Z_t] + \theta_1^2 \text{Var}[Z_{t-1}] = (1 + \theta_1^2)\sigma_Z^2$$

which does not depend on t.

Example: Moving Average

We have

$$\begin{aligned}\gamma_X(t+h,t) &= \mathrm{Cov}[X_{t+h}, X_t] = \mathbb{E}[X_t X_{t+h}] \\ &= \mathbb{E}[(Z_t + \theta_1 Z_{t-1})(Z_{t+h} + \theta_1 Z_{t+h-1})] \\ &= \mathbb{E}[Z_{t+h} Z_t] + \theta_1 \left(\mathbb{E}[Z_{t+h} Z_{t-1}] + \mathbb{E}[Z_{t+h-1} Z_t]\right) \\ &\quad + \theta_1^2 \mathbb{E}[Z_{t-1} Z_{t+h-1}]\end{aligned}$$

Now,

$$\mathbb{E}[Z_j Z_k] = \begin{cases} \sigma_Z^2 & j = k \\ 0 & j \neq k \end{cases}.$$

Example: Moving Average

Hence

$$\gamma_X(t+h,t) = \begin{cases} \sigma_Z^2(1+\theta_1^2) & h=0 \\ \sigma_Z^2\theta_1 & h=\pm 1 \\ 0 & \text{otherwise} \end{cases}.$$

$$\rho_X(t+h,t) = \begin{cases} 1 & h=0 \\ \theta_1/(1+\theta_1^2) & h=\pm 1 \\ 0 & \text{otherwise} \end{cases}$$

These functions do not depend on t, so $\{X_t\}$ is a stationary process

$$\{X_t\} \sim \text{MA}(1)$$

Example: Autoregression

Let $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and define process $\{X_t\}$ by

$$X_t = \phi_1 X_{t-1} + Z_t \qquad t \in \mathbb{Z}$$

where ϕ_1 is a real-valued parameter with $|\phi_1| < 1$.

Assume for the moment that $\{X_t\}$ is stationary, and that

$$\text{Cov}[Z_t, X_s] = 0 \qquad t > s.$$

that is, future values of the Z series are uncorrelated with past values of the X series.

1.4 Stationarity

Example: Autoregression

Then

$$\mathbb{E}[X_t] = \phi_1 \mathbb{E}[X_{t-1}] + \mathbb{E}[Z_t] = \phi_1 \mathbb{E}[X_{t-1}]$$

so therefore by the stationarity assumption $\mathbb{E}[X_t] = 0$.

1.4 Stationarity

Example: Autoregression

We have for all $h > 0$

$$\begin{aligned}
\gamma_X(h) &= \mathrm{Cov}[X_{t+h}, X_t] = \mathrm{Cov}[X_{t-h}, X_t] \\
&= \mathrm{Cov}[\phi_1 X_{t-1}, X_{t-h}] + \mathrm{Cov}[Z_t, X_{t-h}] \\
&= \phi_1 \mathrm{Cov}[X_{t-1}, X_{t-h}] + 0 \\
&= \phi_1 \gamma_X(h-1) \\
&= \phi_1^2 \gamma_X(h-2) \\
&\vdots \\
&= \phi_1^h \gamma_X(0)
\end{aligned}$$

by recursion.

Example: Autoregression

Note that

$$\mathbb{E}[X_t X_{t-h}] = \mathbb{E}[X_{t+h} X_t]$$

by the stationarity assumption. Therefore

$$\gamma_X(-h) = \gamma_X(h)$$

and so

$$\rho_X(h) = \phi_1^{|h|} \qquad h = 0, \pm 1, \pm 2$$

that is, a geometrically decaying ACF.

Example: Autoregression

Note also that

$$\begin{aligned}\gamma_X(0) &= \text{Cov}[X_t, X_t] \\ &= \text{Cov}[(\phi_1 X_{t-1} + Z_t)(\phi_1 X_{t-1} + Z_t)] \\ &= \phi_1^2 \text{Cov}[X_{t-1}, X_{t-1}] + \text{Cov}[Z_t, Z_t] \\ &= \phi_1^2 \gamma_X(0) + \sigma_Z^2\end{aligned}$$

so therefore

$$\gamma_X(0) = \frac{\sigma_Z^2}{1 - \phi_1^2}$$

$\{X_t\}$ is the autoregressive process of order 1: $\{X_t\} \sim \text{AR}(1)$.

Example: Autoregression

From the definition, we have

$$
\begin{aligned}
X_t &= \phi_1 X_{t-1} + Z_t \\
&= \phi_1(\phi_1 X_{t-2} + Z_{t-1}) + Z_t = \phi_1^2 X_{t-2} + (Z_t + \phi_1 Z_{t-1}) \\
&= \phi_1^2(\phi_1 X_{t-3} + Z_{t-2}) + (Z_t + \phi_1 Z_{t-1}) \\
&= \phi_1^3 X_{t-3} + (Z_t + \phi_1 Z_{t-1} + \phi_1^2 Z_{t-2}) \\
&\vdots \\
&= \phi_1^h X_{t-h} + \sum_{j=0}^{h-1} \phi_1^j Z_{t-j}
\end{aligned}
$$

Example: Autoregression

If we allow $h \longrightarrow \infty$, provided $\left|\phi_1\right| < 1$, we can continue the recursion indefinitely to obtain

$$X_t = \sum_{j=0}^{\infty} \phi_1^j Z_{t-j}$$

as the first term converges to zero.

1.4 Stationarity

Note: when we write

$$X_t = Z_t + \theta_1 Z_{t-1}$$

or

$$X_t = \phi_1 X_{t-1} + Z_t$$

we need to be precise about the meaning of the symbol '='.

1.4 Stationarity

For random variables we have (for example)

- equality in distribution

$$X_t \overset{d}{=} Y_t$$

(i.e. the distribution of X_t is the same as the distribution of Y_t)

- equality in probability

$$X_t \overset{p}{=} Y_t$$

(i.e. $\Pr\left[|X_t - Y_t| < \epsilon\right] = 1$ for all $\epsilon > 0$)

1.4 Stationarity

- equality with probability 1 (almost sure equality)

$$X_t \overset{a.s.}{=} Y_t$$

(i.e. For each ω, $\Pr[|X_t(\omega) - Y_t(\omega)| < \epsilon] = 1$ for all $\epsilon > 0$.)

- mean-square equality

$$X_t \overset{m.s.}{=} Y_t$$

(i.e. $\mathbb{E}[(X_t - Y_t)^2] = 0$)

1.4 Stationarity

In the case of an explicit formula, we can interpret the equality as a definition

$$X_t = Z_t + \theta_1 Z_{t-1}$$

but for an implicit formula

$$X_t = \phi_1 X_{t-1} + Z_t$$

we may need to use the other definitions.

1.4 Stationarity

For a stationary process, moment estimators can be used to estimate the mean/covariance structure. The estimates are

$$\widehat{\mu} = \overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n}(x_t - \widehat{\mu})(x_{t+|h|} - \widehat{\mu}) \qquad -n < h < n$$

These are consistent estimators of the two functions. The $n \times n$ matrix estimate

$$\widehat{\mathbf{\Gamma}}_n = [\widehat{\gamma}(i-j)]_{ij}$$

is non-negative definite.

1.5 Trends and Seasonality

The models from the previous section are purely stochastic. It is also possible to incorporate deterministic components.

We focus on

- trends,
- seasonality.

1.5 Trends and Seasonality

Trends: Suppose that for each t,

$$X_t = m_t + E_t$$

where

- m_t is a deterministic function of t,
- E_t is a purely stochastic, zero mean time series.

For example, might have

$$m_t = a_0 + a_1 t$$

that is, a linear trend in time. Constants a_0, a_1 are in general unknown and must be estimated.

1.5 Trends and Seasonality

Ordinary Least-Squares (OLS): estimate a_0, a_1 by minimizing the sum of squared errors

$$\hat{a} = \arg\min \left\{ \sum_{t=1}^{T} (x_t - a_0 - a_1 t)^2 \right\}.$$

Provided $\mathbb{E}[E_t] = 0$, OLS estimators are consistent estimators of the true parameters (if the trend truly is linear)

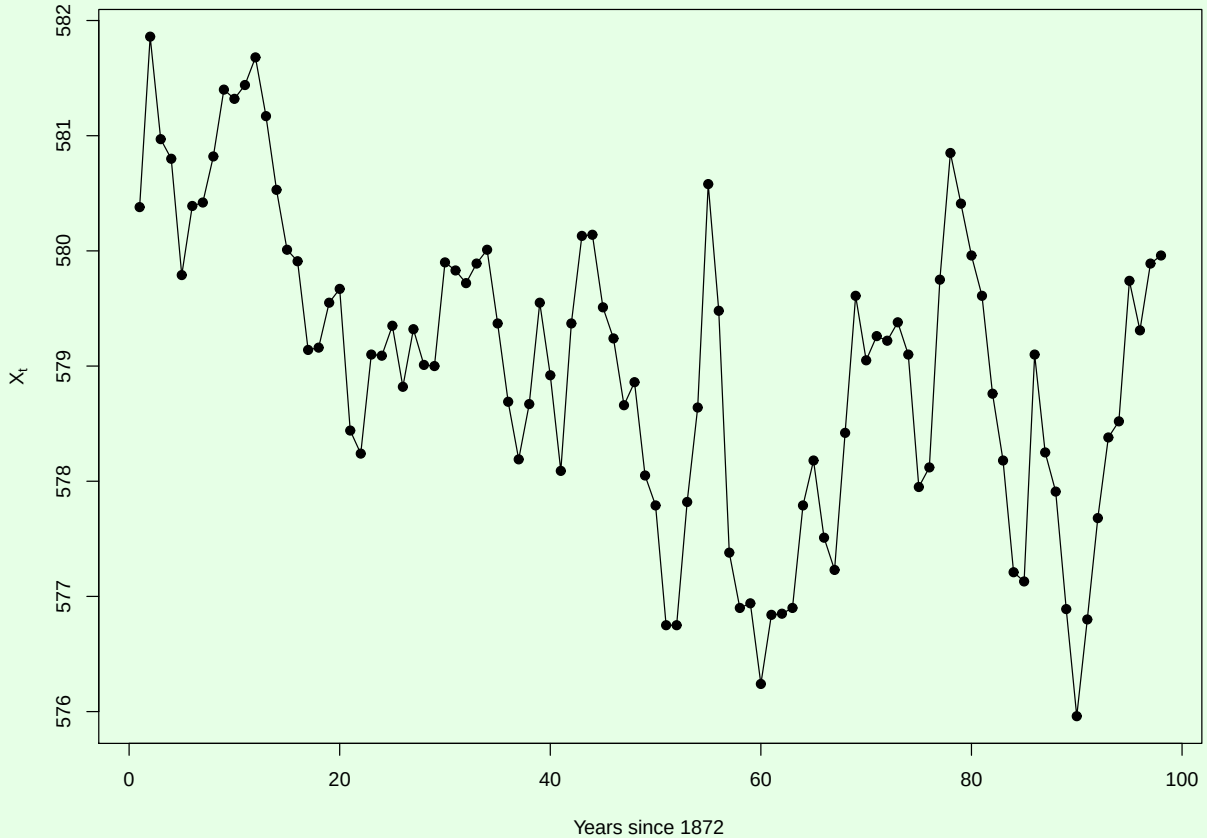
$$\hat{a} \xrightarrow{p} a$$

as $n \longrightarrow \infty$. OLS is easily applied for many types of deterministic trend (e.g. polynomial trend).

1.5 Trends and Seasonality

Example: Lake Huron data

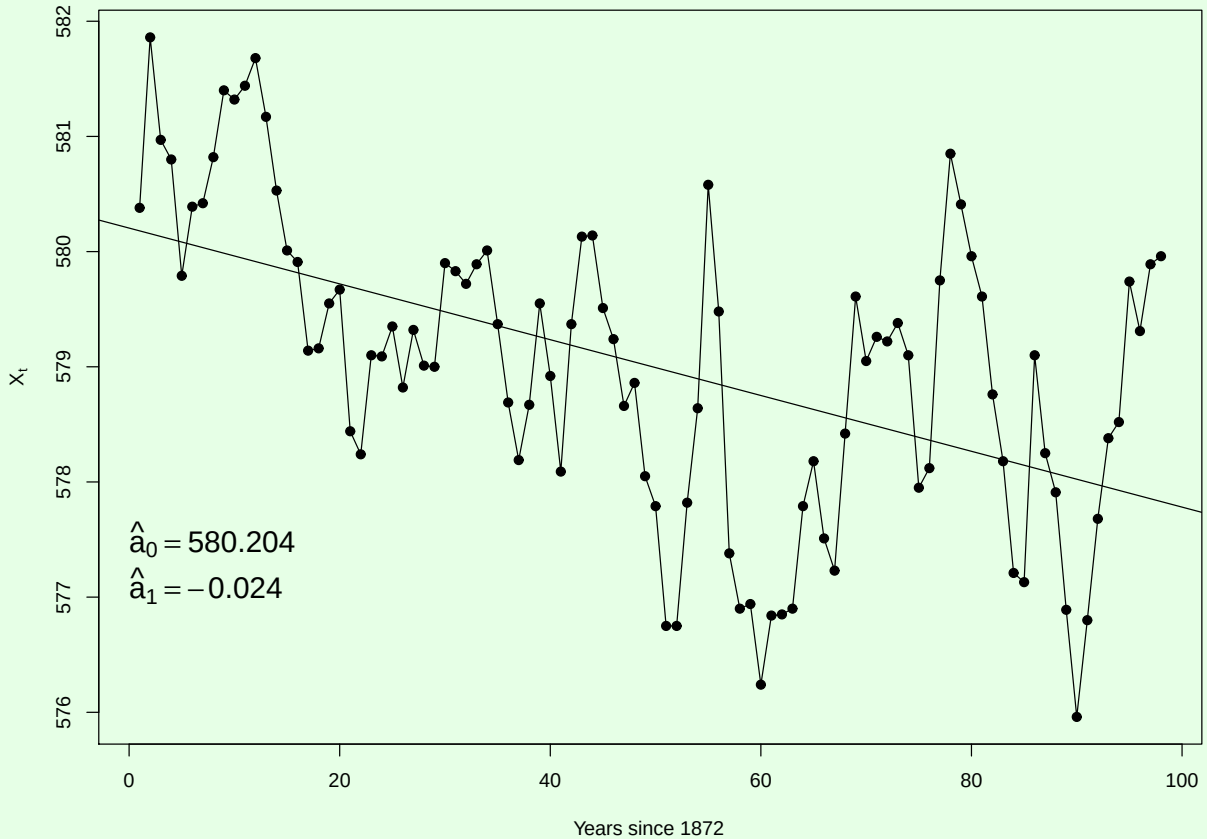
Annual level (in ft) of Lake Huron 1875-1972 ($T = 98$).



1.5 Trends and Seasonality

Example: Lake Huron data

Fitted linear trend: $\hat{a}_0 = 580.204$, $\hat{a}_1 = -0.024$.



1.5 Trends and Seasonality

Example: Lake Huron data

After fitting the linear trend model

$$X_t = a_0 + a_1 t + E_t$$

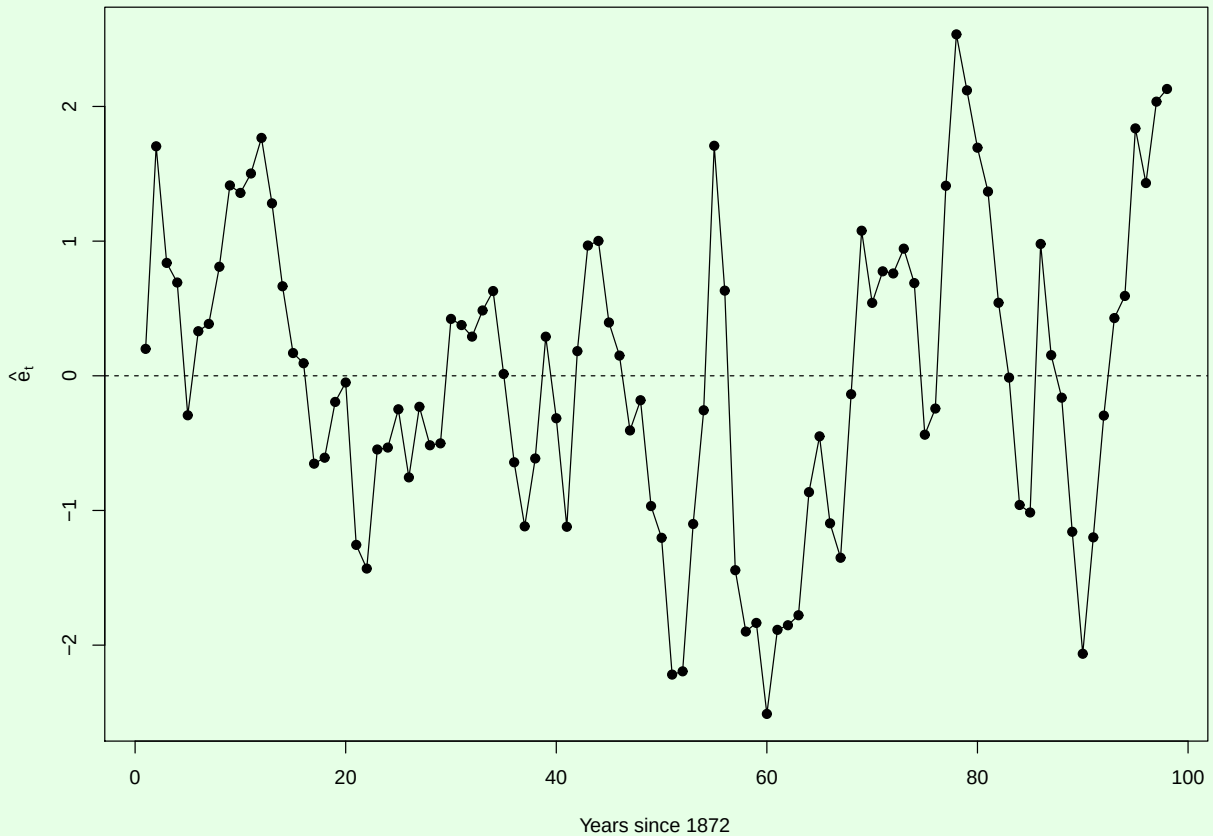
by OLS to obtain $\hat{a}_0, \hat{a}_1$, we may examine the residual series

$$\hat{e}_t = x_t - \hat{a}_0 - \hat{a}_1 t$$

1.5 Trends and Seasonality

Example: Lake Huron data

Residual series: $\hat{e}_t = x_t - \hat{a}_0 - \hat{a}_1 t.$



Example: Lake Huron data

The residual series is approximately zero mean. However, there is some structure:

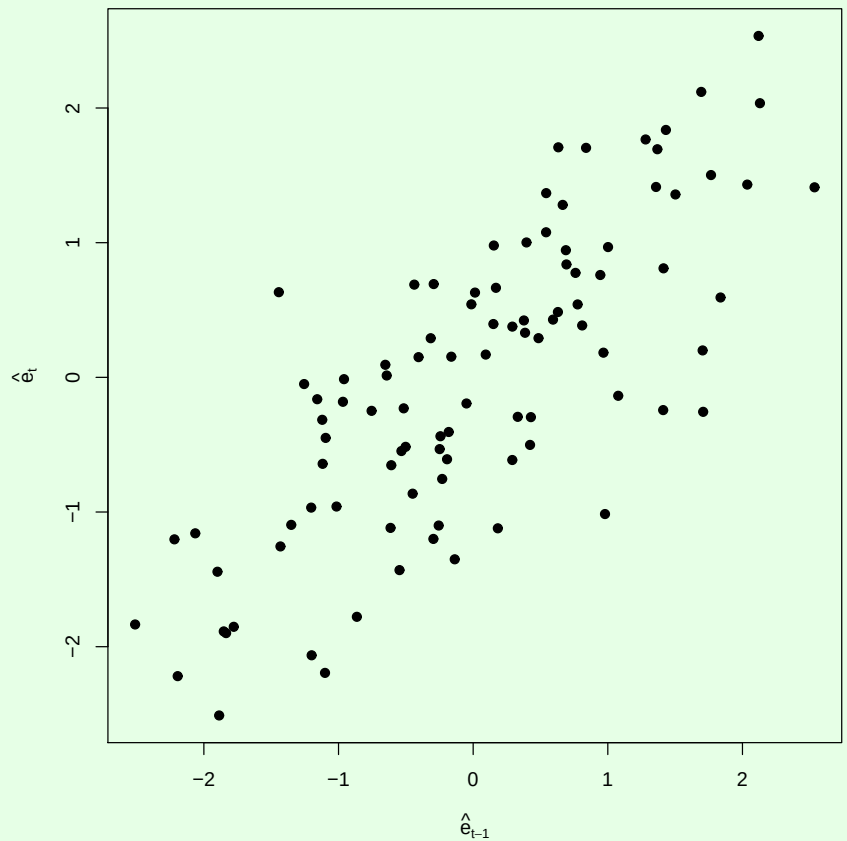
- the variance is not constant over time ?
- successive residuals $(\hat{e}_{t-1}, \hat{e}_t)$ are positively correlated. The sample correlation is

$$\text{corr}[\hat{e}_{t-1}, \hat{e}_t] \approx 0.775$$

That is, the residual series **does not** appear to be a realization of an IID process.

1.5 Trends and Seasonality

Example: Lake Huron data: $\widehat{e}_t$ vs $\widehat{e}_{t-1}$.



Example: Lake Huron data

The positive dependence is potentially useful for forecasting; if

$$\hat{e}_{93}, \hat{e}_{94}, \ldots, \hat{e}_{98}$$

are positive, it is likely that

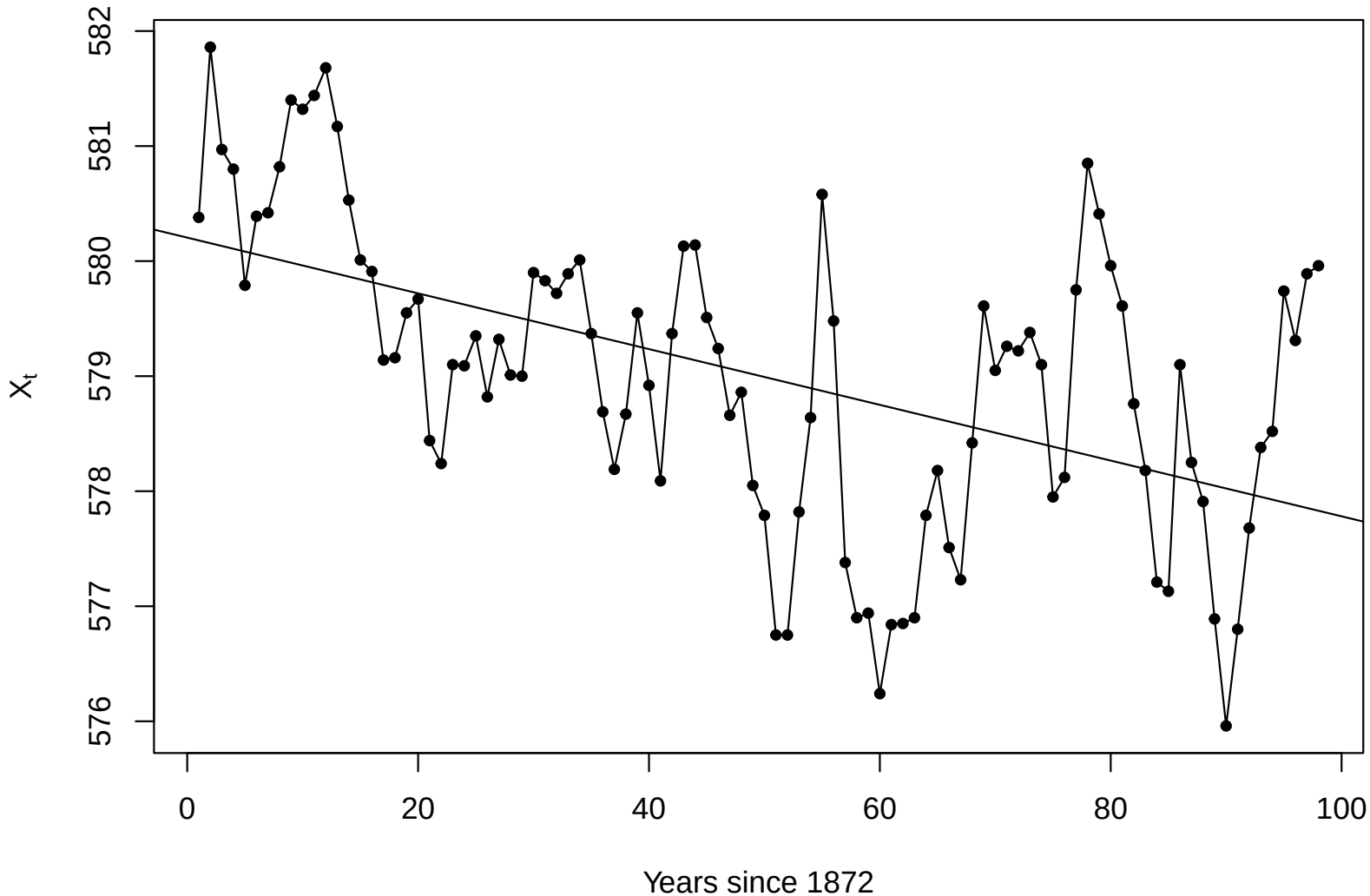
$$\hat{e}_{99}$$

will also be positive. That is, conditional on the available data, the prediction for the next data point X_{99} will be different if the residuals apparently exhibit correlation.

R Code

```r
Huron<-scan("https://www.math.mcgill.ca//dstephens//545//Data//huron.txt")
Huron<-Huron+570
n.H<-length(Huron); t.H<-c(1:n.H); x<-c(1875:1972)
par(mar=c(4,4,2,2))
plot(t.H,Huron,type="l",
     xlab="Years since 1872",ylab=expression(X[t]))
points(t.H,Huron,pch=19,cex=0.7)
abline(lm(Huron~t.H))
```

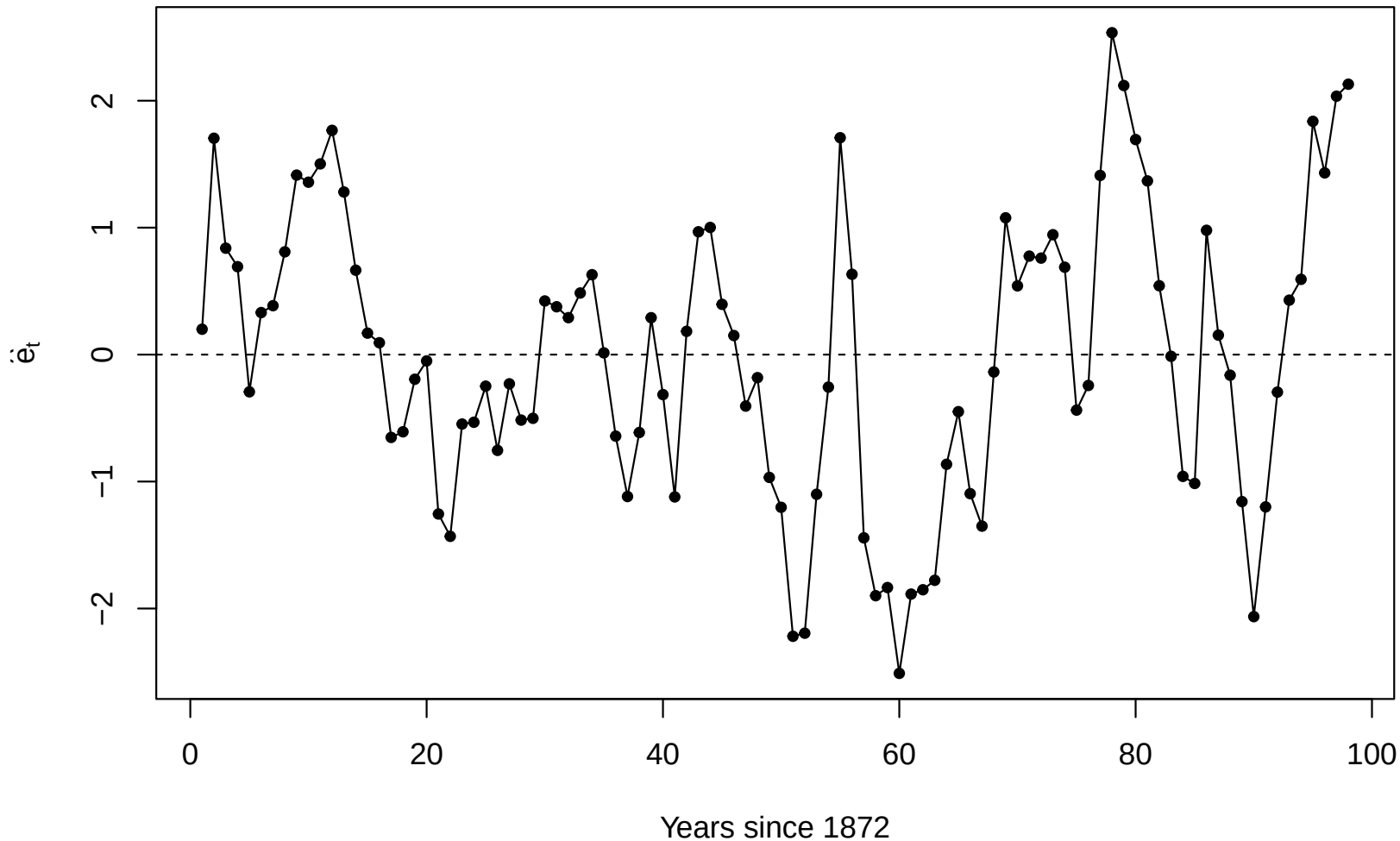
R Code



R Code

```
a0<-lm(Huron~t.H)$coeff[1]; a1<-lm(Huron~t.H)$coeff[2]
mt.H<-a0+a1*t.H; et.H<-Huron-mt.H
par(mar=c(4,4,2,2))
plot(t.H,et.H,type="l",xlab="Years since 1872",
                       ylab=expression(hat(e)[t]))
points(t.H,et.H,pch=19,cex=0.7); abline(h=0,lty=2)
```

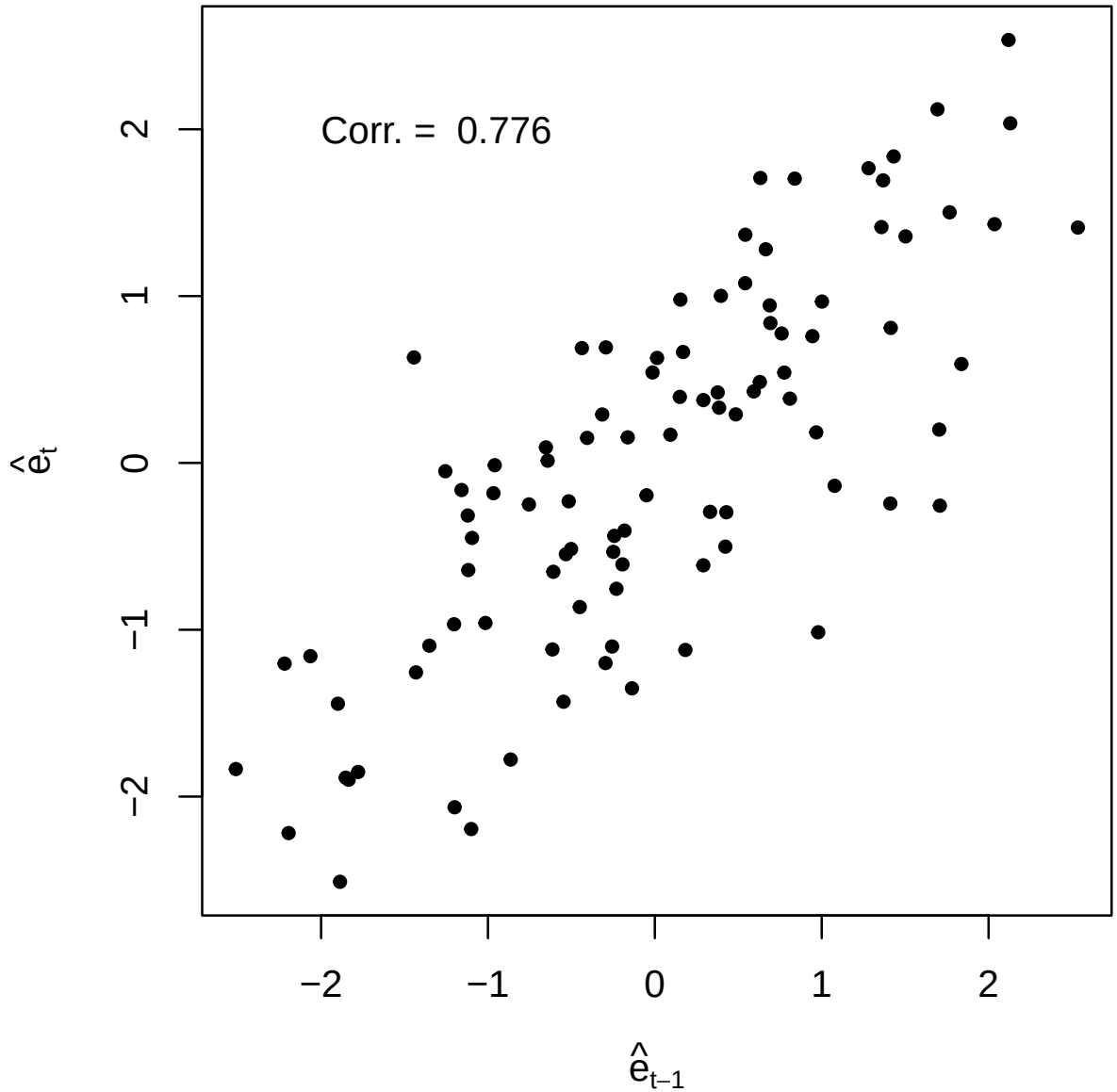
R Code



R Code

```
y1<-et.H[2:98];y2<-et.H[1:97];cor(y1,y2)
par(mar=c(5,4,2,2),pty='s')
plot(y1,y2,xlab=expression(hat(e)[t-1]),
           ylab=expression(hat(e)[t]),pch=19,cex=0.7)
```

R Code



Example: Lake Huron data

Recall that the model

$$X_t = m_t + E_t$$

was fitted using a linear trend $m_t = a_0 + a_1 t$ leaving residuals that were positively correlated (see plot on page 58). We may also model

$$E_t \sim \text{AR}(1),$$

that is, assume

$$E_t = \phi_1 E_{t-1} + Z_t$$

and estimate ϕ_1 using OLS from the fitted residuals e_t. This assumes that $\{Z_t\} \sim WN(0, \sigma_Z^2)$.

Example: Lake Huron data

This procedure yields $\widehat{\phi}_1 = 0.791$. By inspecting the fitted values, we may construct the residual quantities

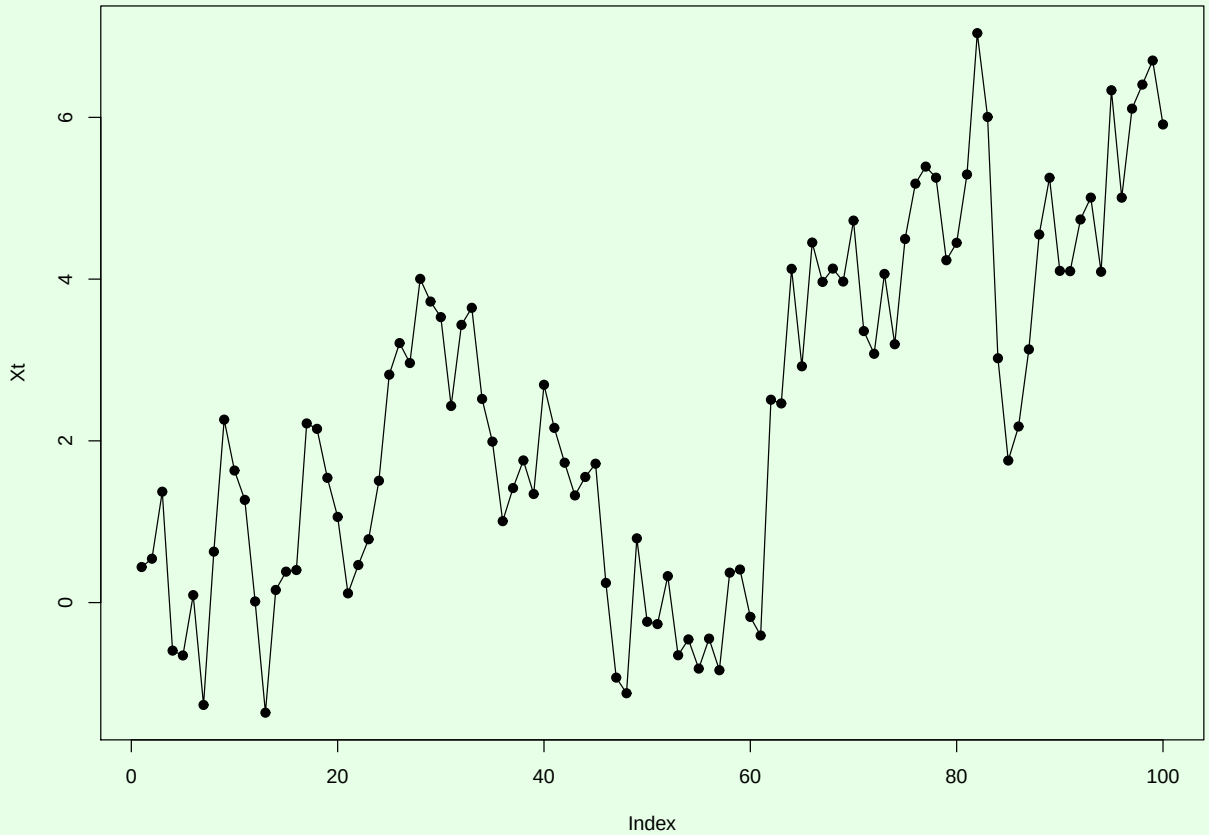
$$\widehat{z}_t = e_t - \widehat{\phi}_1 e_{t-1}$$

we may estimate σ_Z^2 via the sample variance of the $\widehat{z}_t$. This yields $\widehat{\sigma}^2 = 0.502$.

If the AR(1) model were correct, then the residual quantities $\widehat{z}_t$ should resemble a realization of a white noise process; however, they do not - there is still positive correlation between successive $\widehat{z}_t$s.

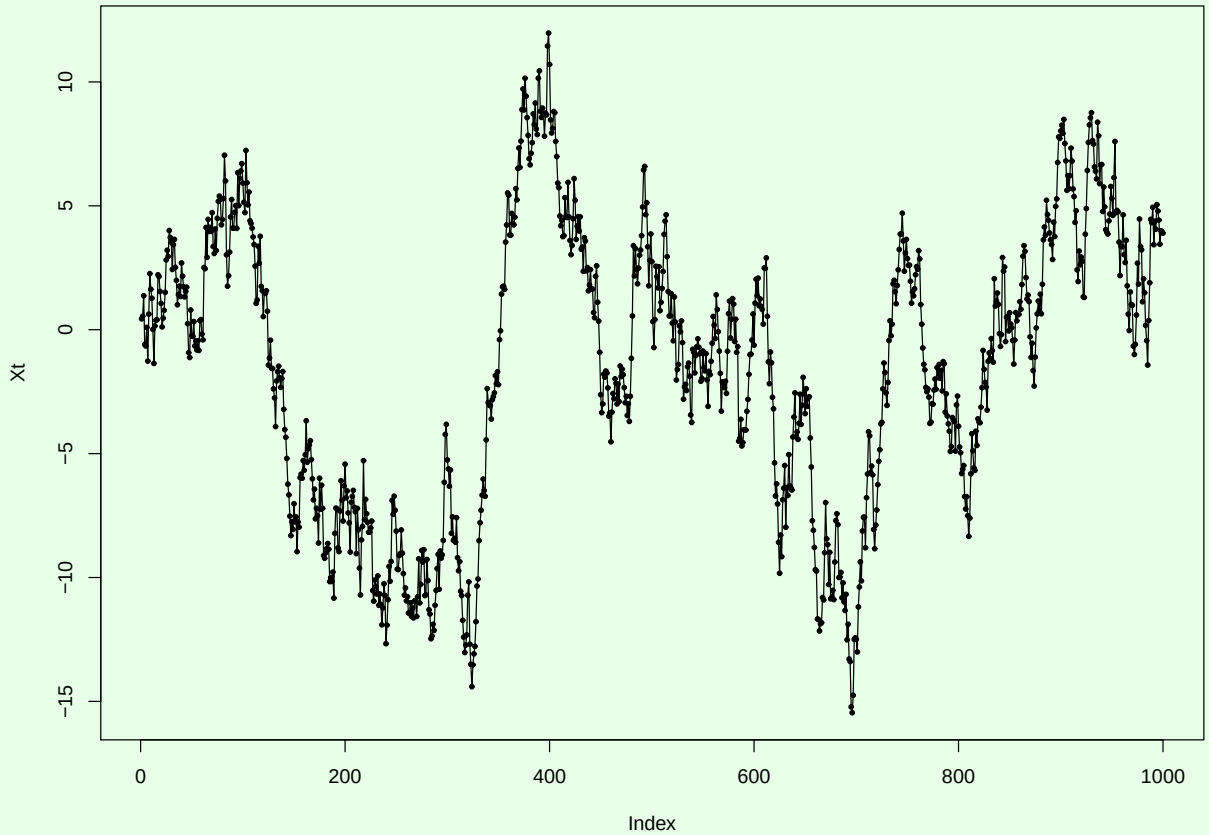
Example: Trend ?

100 observations



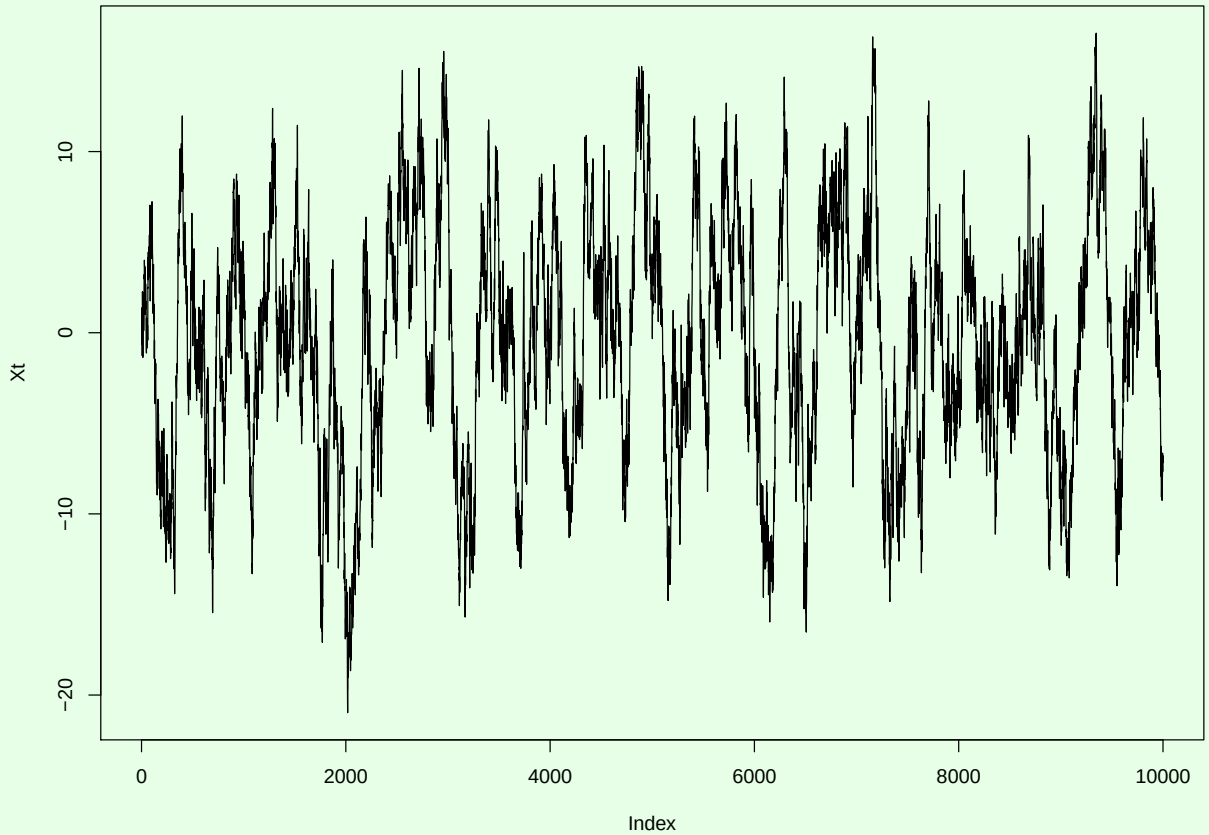
Example: Trend ?

1000 observations



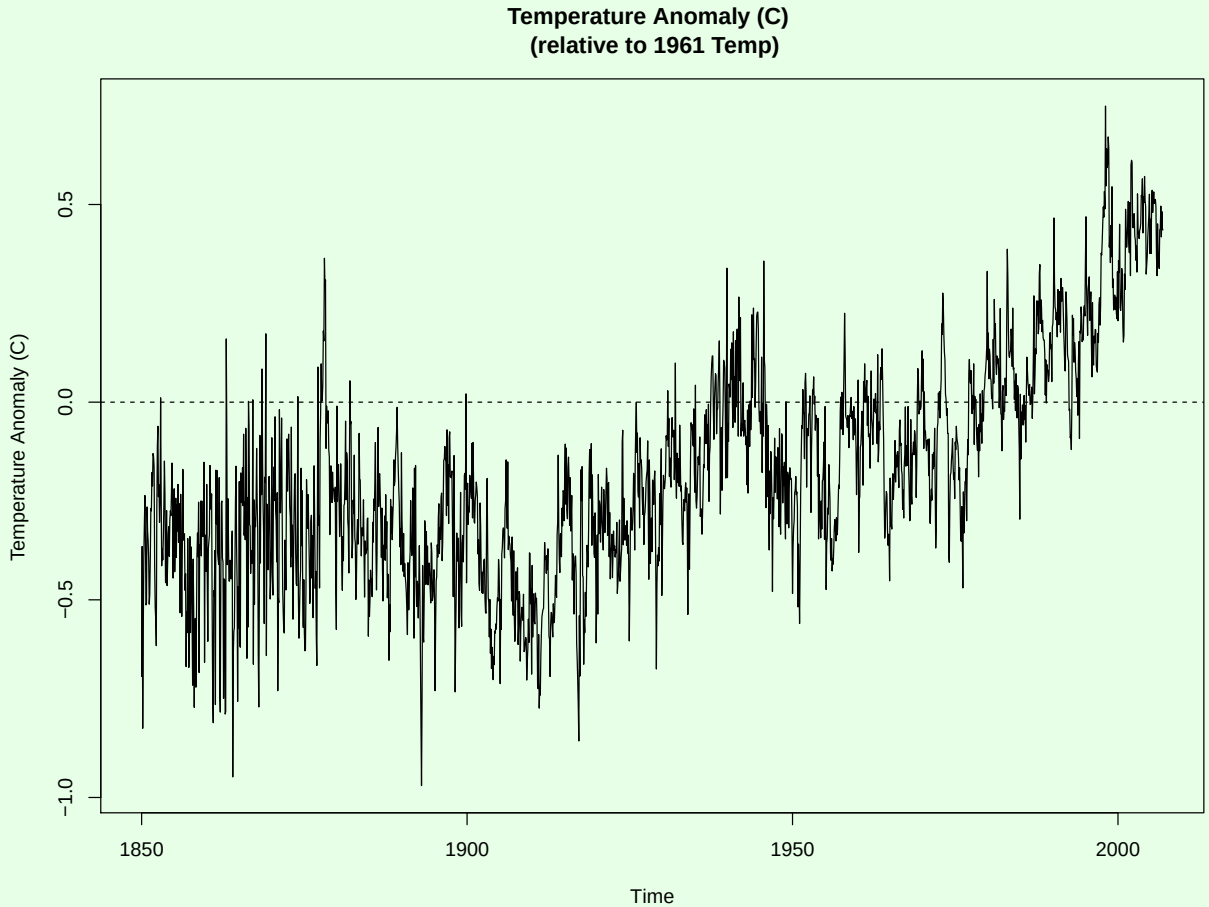
Example: Trend ?

10000 observations



1.5 Trends and Seasonality

Example: Temperature Anomaly: 1850-2006



1.5 Trends and Seasonality

Seasonality: Many time series are influenced by seasonally-varying factors

- calendar
- climate
- economic cycles

1.5 Trends and Seasonality

A deterministic seasonal model can be constructed. Suppose that for each t,

$$X_t = s_t + E_t$$

where s_t is a periodic function with period d say, so that

$$s_{t-d} = s_t$$

for all t.

1.5 Trends and Seasonality

For example, could use a harmonic regression

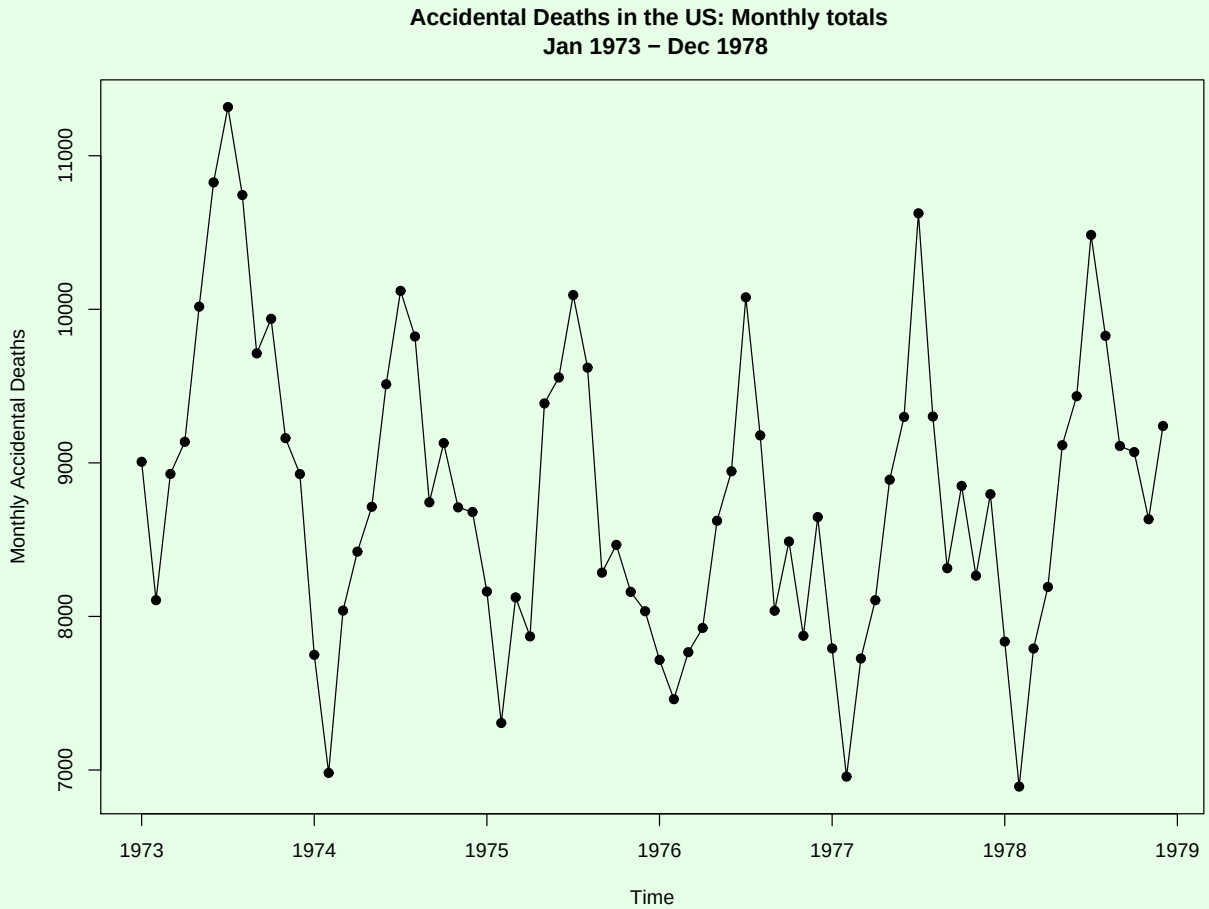
$$s_t = a_0 + \sum_{j=1}^{k} [a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)]$$

where $a_0, a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ are unknown coefficients to be estimated, and

$$\lambda_1, \ldots, \lambda_k$$

are known constants that define the periodic nature of the function.

Example: Accidental Deaths in the US



1.5 Trends and Seasonality

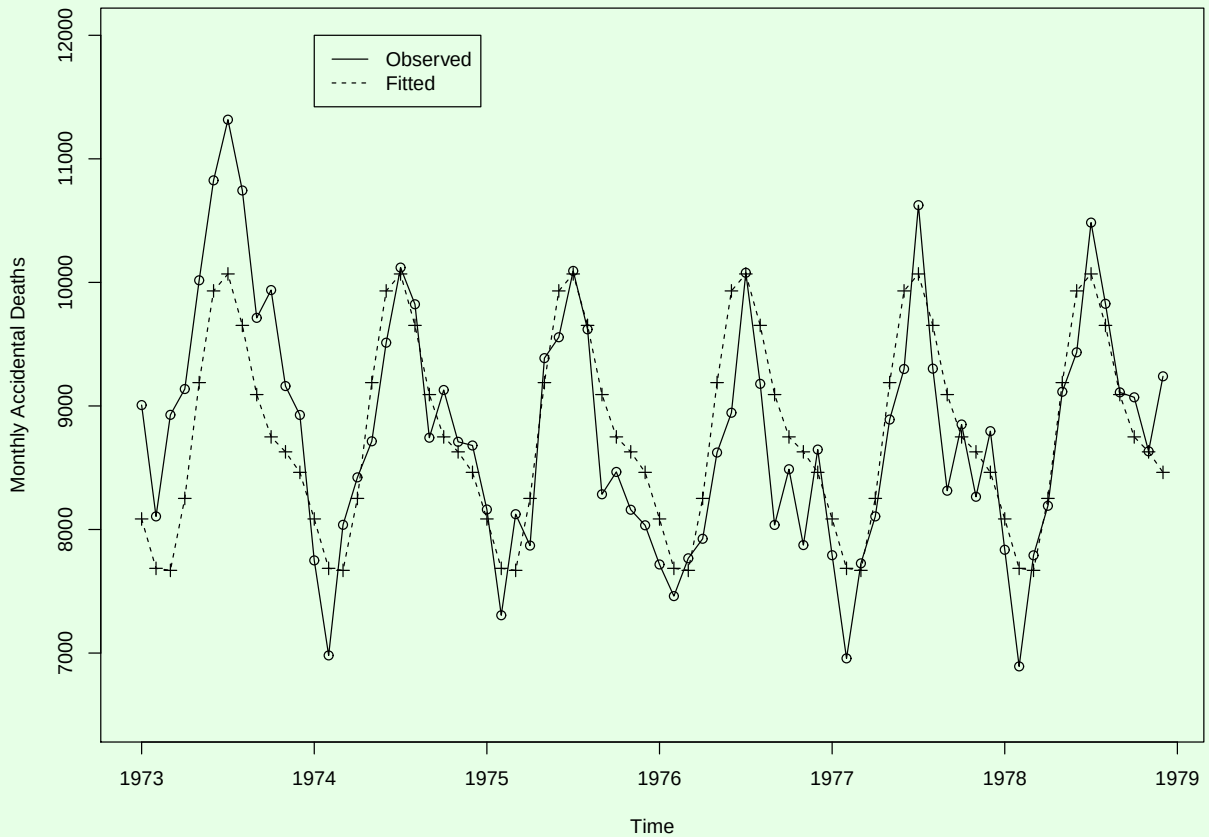
Example: Accidental Deaths in the US

For this data set $T = 72$, for monthly data. To fit a model with $k = 2$

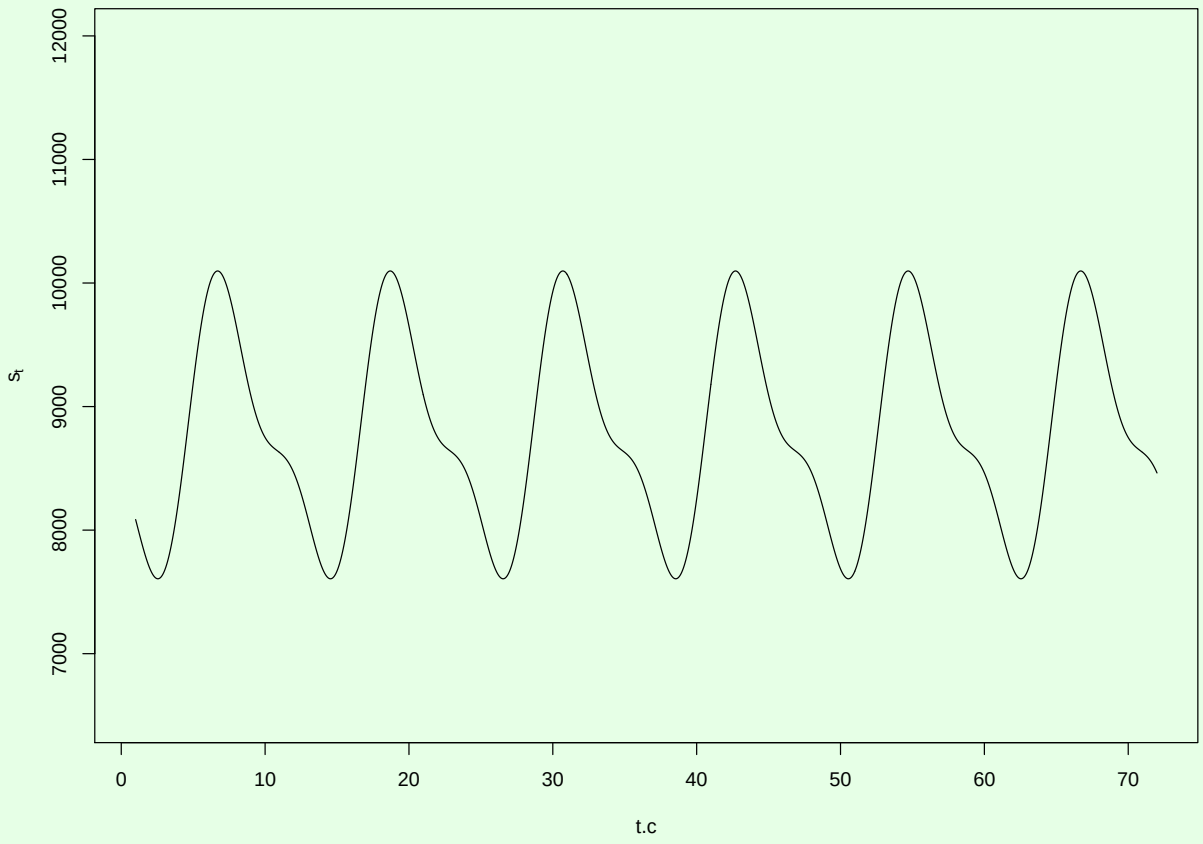
- 12 month cycle
- 6 month cycle

set $\lambda_1 = 2\pi/12$, $\lambda_2 = 2\pi/6$. Then fit using OLS.

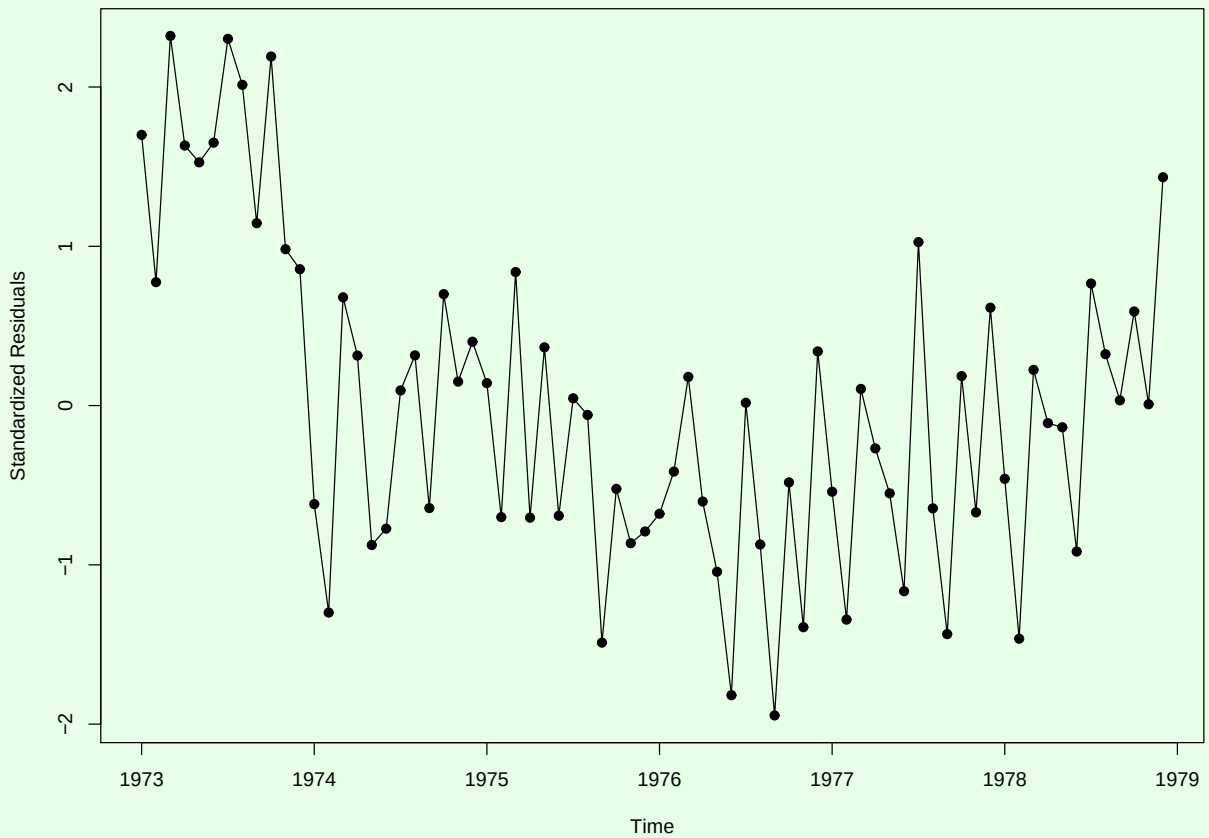
Example: Accidental Deaths in the US



Example: Accidental Deaths in the US: Harmonics



Example: Accidental Deaths in the US: Residuals



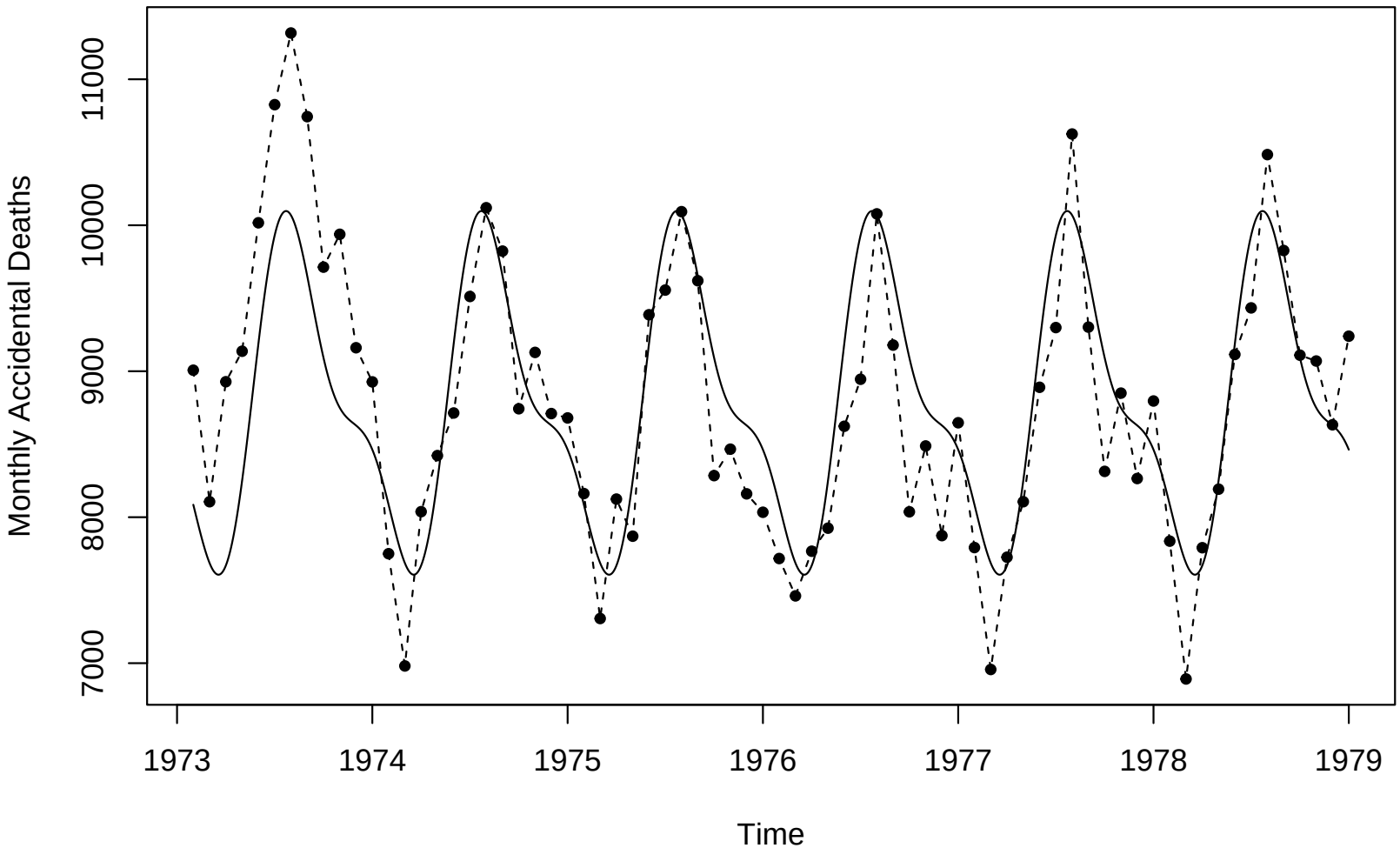
R Code

```
library(datasets)
t.v<-c(1:72)
C1<-cos(2*pi*t.v/12); S1<-sin(2*pi*t.v/12)
C2<-cos(2*pi*t.v/6); S2<-sin(2*pi*t.v/6)
USA.ACC<-data.frame(as.numeric(USAccDeaths),S1,C1,S2,C2)
USA.ACC.fit<-lm(USAccDeaths~S1+C1+S2+C2, data=USA.ACC)$fitted
USA.ACC.coef<-lm(USAccDeaths~S1+C1+S2+C2,data=USA.ACC)$coef
t.c<-seq(from=1,to=72,by=0.01)
C1.c<-cos(2*pi*t.c/12); S1.c<-sin(2*pi*t.c/12)
C2.c<-cos(2*pi*t.c/6); S2.c<-sin(2*pi*t.c/6)
y.c<-USA.ACC.coef[1]+USA.ACC.coef[2]*S1.c+
     USA.ACC.coef[3]*C1.c +USA.ACC.coef[4]*S2.c+
     USA.ACC.coef[5]*C2.c
```

R Code

```
par(mar=c(4,4,2,2))
plot(1:72,as.numeric(USAccDeaths),pch=19,cex=0.7,xaxt='n',
    xlab='Time',ylab='Monthly Accidental Deaths')
lines(1:72,as.numeric(USAccDeaths),lty=2)
lines(t.c,y.c)
axis(1,at=12*c(0:6),c(1973:1979))
```

R Code


Monthly Accidental Deaths
7000
8000
9000
10000
11000
1973
1974
1975
1976
1977
1978
1979
Time

1.6 A Decomposition Strategy

Recap: We seek to decompose an observed series $\{Y_t\}$ as

$$Y_t = m_t + s_t + X_t$$

where

- m_t is a deterministic trend
- s_t is a seasonal component with period d
- X_t is a zero mean stationary process.

It may be necessary to replace the global model for trend (linear, polynomial etc.) by a local model

1.6 A Decomposition Strategy

Smoothing by averaging: Suppose

$$\hat{m}_t = \frac{1}{2q+1} \sum_{i=-q}^{q} x_{t+i}$$

with $x_s = 1$ for $s < 1$ and $s > T$. More generally,

$$\hat{m}_t = \sum_{i=-q}^{q} a_i x_{t+i}$$

for constants $a_{-q}, \ldots, a_q$.

This is a form of low-pass filter.

1.6 A Decomposition Strategy

Exponential Smoothing: Suppose $\hat{m}_1 = x_1$, and for $t \geqslant 2$

$$\hat{m}_t = \alpha x_t + (1 - \alpha)\hat{m}_{t-1}$$

for $0 < \alpha < 1$. It follows that

$$\hat{m}_t = (1 - \alpha)^{t-1} x_1 + \sum_{j=1}^{t-2} \alpha(1 - \alpha)^j x_{t-j} + \alpha x_t$$

If α is near one, there is strong dependence on recent x values.

1.6 A Decomposition Strategy

General Strategy: For series $\{Y_t\}$

$$Y_t = m_t + s_t + X_t$$

where s_t has period d.

1. If d is even, set $q = d/2$, else set $q = (d-1)/2$.
2. Use moving average for detrending respecting the seasonality

$$\widehat{m}_t = \begin{cases} \dfrac{y_{t-q} + y_{t+q}}{2d} + \dfrac{1}{d} \displaystyle\sum_{j=-q+1}^{q-1} y_{t+j} & d \text{ even} \\ \dfrac{1}{d} \displaystyle\sum_{j=-q}^{q} y_{t+j} & d \text{ odd} \end{cases}$$

3. Remove seasonality: for $k = 1, \ldots, d$, compute

$$w_k = \frac{1}{n_d} \sum_j (y_{k+jd} - \hat{m}_{k+jd})$$

where n_d is the number of terms in the sum (essentially the number of cycles in the series). For $k = 1, \ldots, d$, set

$$\hat{s}_k = w_k - \frac{1}{d} \sum_{j=1}^{d} w_j$$

and set $\hat{s}_k = \hat{s}_{k-d}$ for $k = d+1, d+2, \ldots, T$. The deseasonalized data are $y_t^\star = y_t - \hat{s}_t$.

1.6 A Decomposition Strategy

4. Recompute the deseasonalized trend $\widehat{m}_t^\star$ from the deseasonalized data $y_t^\star$ using the procedure in Step 2.
5. Compute the residual series

$$\widehat{x}_t = y_t - \widehat{m}_t^\star - \widehat{s}_t.$$

The resulting process $\widehat{x}_t$ should be a zero mean, but potentially autocorrelated process.

1.7 Differencing

Differencing: Another method for removing non-stationarity is differencing. The lag-1 difference operator, ∇, acts on X_t as follows

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t$$

where B is the backshift operator

$$BX_t = X_{t-1}.$$

Note that

$$B^2 X_t = B(BX_t) = BX_{t-1} = X_{t-2}$$

etc.

1.7 Differencing

Similarly

$$\begin{aligned}
\nabla^2 X_t &= (1-B)^2 X_t \\
&= (1-B)((1-B)X_t) \\
&= (1-B)(X_t - X_{t-1}) \\
&= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\
&= X_t - 2X_{t-1} + X_{t-2}
\end{aligned}$$

Note that k times differencing removes a polynomial trend of order k.

1.7 Differencing

Seasonal Differencing: The lag-d difference operator, ∇_d, acts on X_t as follows

$$\nabla_d X_t = (1 - B^d) X_t = X_t - X_{t-d}$$

This operator removes a seasonality with period d.

∇ and ∇_d can be applied consecutively; they commute

$$\nabla \nabla_d X_t = \nabla_d \nabla X_t$$

1.7 Differencing

Elementary algebraic properties:

(i)

$$B^j X_t = X_{t-j} = BX_{t-j+1} = B^2 X_{t-j+2} = \cdots$$

(ii) $\nabla^j X_t = \nabla(\nabla^{j-1} X_t)$. So for example,

$$\begin{aligned}
\nabla^2 X_t &= (1-B)(1-B)X_t \\
&= (1 - 2B + B^2)X_t \\
&= X_t - 2X_{t-1} + X_{t-2} \\
&= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2})
\end{aligned}$$

(iii) If $X_t = m_t + Y_t$, applying the first-difference operator to X_t, then

$$\nabla X_t = X_t - X_{t-1} = (m_t - m_{t-1}) + (Y_t - Y_{t-1})$$

so if $m_t = \beta_0 + \beta_1 t$, then $m_t - m_{t-1} = \beta_1$.

That is, ∇ removes a linear trend.

1.7 Differencing

The stochastic term

$$Y_t^* = (Y_t - Y_{t-1})$$

is obtained by differencing.

This is different from our first approach to detrending, which used a linear parametric formulation for m_t.

- If Y_t is white-noise, then Y_t^* is no longer white noise, but MA(1).

1.7 Differencing

(iv) To remove a polynomial trend of order k, one may apply k^{th} order differencing, that is we look at $\nabla^k X_t$.

Note that if $\{Y_t\} \sim \text{WN}(0, \sigma_Y^2)$, then $\nabla^k Y_t$ is not white-noise, but is still stationary.

1.7 Differencing

(v) Commutativity

$$\nabla\nabla_d X_t = (1-B)(1-B^d)X_t = (1-B^d)(1-B)X_t = \nabla_d \nabla X_t$$

(vi) If

$$X_t = m_t + s_t + Y_t,$$

then

$$\nabla^d X_t = (m_t - m_{t-d}) + (s_t - s_{t-d}) + (Y_t - Y_{t-d})$$

and

$$s_t - s_{t-d} = 0.$$

That is, ∇_d removes a seasonality with period d.

Chapter 2
Stationary processes

2.1 Stationary Processes

The objective of the previous section was to remove (systematic) deterministic components from an observed series to leave only the stochastic part. In this section, we study the basic properties of stationary processes: such processes are inherently stable (in the long run), and form natural models for the stochastic component of observed series.

In the main, we will focus on weakly stationary processes, where the mean and covariance structure is stable over time; in general we will not make distributional assumptions.

Initially, we focus on forecasting, and demonstrate the utility of utilizing the covariance structure of the process.

2.2 Properties of Stationary Processes

Consider first a stationary Gaussian process $\{X_t\}$, where for all t and n,

$$(X_{t+1}, \ldots, X_{t+n})$$

is jointly normally distributed, with $\mathbb{E}[X_t] = \mu$, $\text{Var}[X_t] = \sigma_X^2$, and for $s > t$.

$$\text{Cov}[X_t, X_s] = \gamma_X(s-t) \qquad \text{Corr}[X_t, X_s] = \rho_X(s-t) = \frac{\gamma_X(s-t)}{\sigma_X^2}$$

Then it follows that

$$X_{n+h} | X_n = x_n \sim \mathcal{N}(m_h(x_n), v_h(x_n))$$

where

$$\begin{aligned} m_h(x) &= \mu + \rho_X(h)(x - \mu) \\ v_h(x) &= \sigma_X^2(1 - \rho_X^2(h)) \end{aligned}$$

2.2 Properties of Stationary Processes

To define “optimal" forecasting, we need to define a criterion via which to assess the quality of predictions.

We use a minimum mean-square error (minimum MSE) criterion, and attempt to to make the prediction $\widehat{X}_{n+h} = \widehat{x}_{n+h}$ such that

$$\mathbb{E}\left[\left(X_{n+h} - \widehat{x}_{n+h}\right)^2 \middle| X_n = x_n\right]$$

is minimized. Here the expectation is taken with respect to the conditional distribution $X_{n+h} \middle| X_n = x_n$.

This minimization can be carried out analytically to yield

$$\hat{x}_{n+h} = \mathbb{E}[X_{n+h}|X_n = x_n] = m_h(x) = \mu + \rho_X(h)(x_n - \mu)$$

that is, the best prediction (measured in minimum prediction MSE terms) is the conditional expectation. The prediction MSE corresponding to this prediction is

$$v_h(x) = \sigma_X^2(1 - \rho_X^2(h))$$

In this case, we could deduce the appropriate form for the optimal predictor from properties of the Gaussian process.

2.2 Properties of Stationary Processes

Best Linear Prediction: For a non-Gaussian process, we seek the again seek the minimum MSE predictor. We restrict attention to linear predictors, that is, predictors of the form

$$\widehat{X}_{n+h} = l_h(X_n) = aX_n + b$$

for some a and b. Under this restriction, and stationarity, it again transpires that the optimal predictor takes the form

$$\widehat{x}_{n+h} = \mathbb{E}[X_{n+h}|X_n = x_n] = m_h(x) = \mu + \rho_X(h)(x_n - \mu)$$

with the same prediction MSE as before.

2.2 Properties of Stationary Processes

Non-negative definite functions: A scalar function g is non-negative definite if, for all n, and vectors $a = (a_1, \ldots, a_n)$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i g(i-j) a_j \geqslant 0.$$

If the summation is strictly greater than zero, the function is termed positive definite.

If G is the $n \times n$ matrix formed as $G = [g(i-j)]_{ij}$, then the above definition becomes that

$$a^\top G a \geqslant 0.$$

and G is a non-negative/positive definite matrix.

2.2 Properties of Stationary Processes

Theorem:

A function g is the autocovariance function (ACVF) for a stationary stochastic process if and only if g is **even**, that is

$$g(h) = g(-h) \qquad \text{for all } h \in \mathbb{Z}$$

and g is **non-negative definite**.

Note: by elementary properties of covariance, for all h,

$$\begin{aligned} \gamma(0) &\geqslant 0 \\ |\gamma(h)| &\leqslant \gamma(0) \\ \gamma(-h) &= \gamma(h) \end{aligned}$$

Characterizing the ACVF

Proof. Suppose that g is the ACVF of a stationary stochastic process, $\{X_t\}$ say. By definition of covariance, g is an even function, as

$$g(-h) = \mathbb{E}[X_t X_{t-h}] = \mathbb{E}[X_{t+h} X_t] = \mathbb{E}[X_t X_{t+h}] = g(h).$$

Let $a \in \mathbb{R}^n$ be an $n \times 1$ vector, and let $X_{n:1} = (x_n, \ldots, x_1)^\top$. Then $a^\top X_{n:1}$ is a scalar random variable, and

$$\mathrm{Var}[a^\top X_{n:1}] = a^\top \mathrm{Var}[X_{n:1}] a = \sum_{i=1}^{n} \sum_{j=1} a_i g(i-j) a_j \geqslant 0$$

so g is non-negative definite.

Characterizing the ACVF

Now suppose that g is non-negative definite and even. For each $n \geqslant 1$, define the $n \times n$ matrix Σ by

$$[\Sigma]_{ij} = g(i-j)$$

and consider the multivariate Normal distribution with mean zero and variance-covariance matrix Σ. This distribution is the finite dimensional distribution arising from a stationary Gaussian process with the specified covariance structure. $\square$

Note that an autocorrelation function has the same properties, except that $\rho(0) = 1$.

2.2 Properties of Stationary Processes

Example: Simple harmonic process

Suppose U and V are zero mean and uncorrelated random variables with variance 1. Let $\{X_t\}$ be defined by

$$X_t = U\cos(\omega t) + V\sin(\omega t)$$

where ω is a fixed constant.

U and V define random amplitudes.

Example: Simple harmonic process

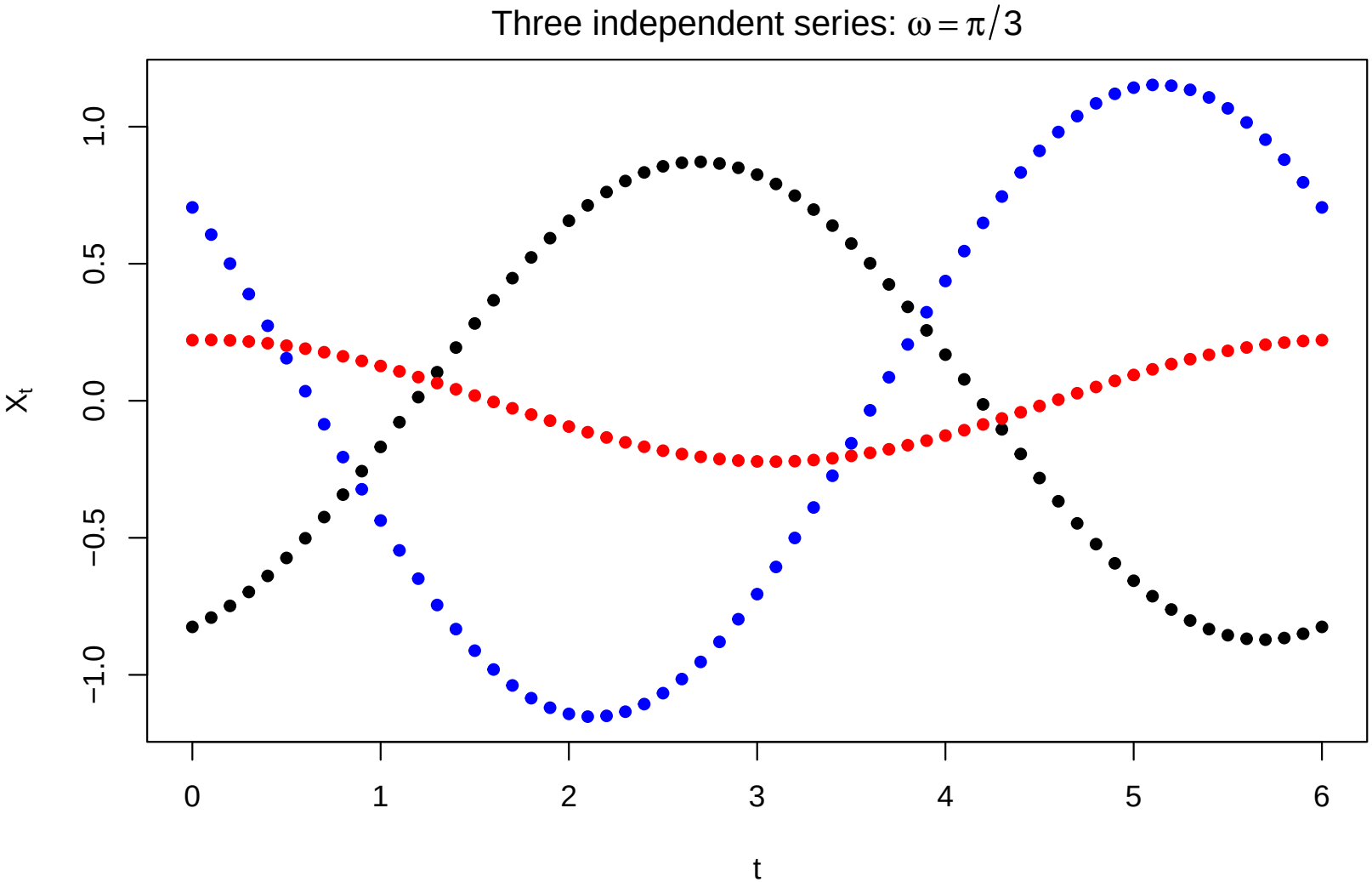
Then $\mathbb{E}[X_t] = 0$ for all t, and

$$\begin{aligned}
\operatorname{Cov}[X_t, X_{t+h}] &= \mathbb{E}[X_t X_{t+h}] \\
&= \mathbb{E}\left[(U\cos(\omega t) + V\sin(\omega t))\right. \\
&\qquad \left.(U\cos(\omega(t+h)) + V\sin(\omega(t+h)))\right] \\
&= \mathbb{E}[U^2]\cos(\omega t)\cos(\omega(t+h)) + \\
&\qquad \mathbb{E}[V^2]\sin(\omega t)\sin(\omega(t+h)) \\
&= \cos(\omega t)\cos(\omega(t+h)) + \sin(\omega t)\sin(\omega(t+h)) \\
&= \cos(\omega h) = \cos(\omega(-h))
\end{aligned}$$

2.2 Properties of Stationary Processes

```
omega<-pi/3
set.seed(2343)
xt<-seq(0,6,by=0.1)
y1<-rnorm(1)*cos(omega*xt)+rnorm(1)*sin(omega*xt)
y2<-rnorm(1)*cos(omega*xt)+rnorm(1)*sin(omega*xt)
y3<-rnorm(1)*cos(omega*xt)+rnorm(1)*sin(omega*xt)
ymin<-min(y1,y2,y3)
ymax<-max(y1,y2,y3)
par(mar=c(4,4,2,2))
plot(xt,y1,xlab='t',ylab=expression(X[t]),
        main=expression(paste('Three independent series: ',omega==pi/3)),
        ylim=range(ymin,ymax),pch=19,cex=0.8)
points(xt,y2,col='red',pch=19,cex=0.8)
points(xt,y3,col='blue',pch=19,cex=0.8)
```

2.2 Properties of Stationary Processes



2.3 Construction of stationary processes

Let $\{Z_t\}$ be an IID process, and let $\{X_t\}$ be defined by

$$X_t = f(Z_t, Z_{t-1}, \ldots, Z_{t-q})$$

for some $q \geqslant 0$. $\{X_t\}$ is a filtered version of $\{Z_t\}$.

- $\{X_t\}$ is strictly stationary as $\{Z_t\}$ is strictly stationary.
- X_t and X_s are independent if $|t - s| > q$; $\{X_t\}$ is termed q-dependent, and $\gamma_X(h) = 0$ if $|h| > q$.

2.3 Construction of stationary processes

Example: Moving average processes

Let $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$ be a white noise process, and let $\{X_t\}$ be defined by

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}$$

for some $q \geqslant 0$, where $\theta_1, \theta_2, \ldots, \theta_q$ are real-valued constants.

$\{X_t\}$ is zero mean, stationary and q-dependent; it is termed a moving average process of order q (MA(q)).

2.4 Linear Processes

$\{X_t\}$ is a linear process if for all t

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

where

- $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$
- $\{\psi_j\}$ is a real sequence with

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

2.4 Linear Processes

Using the backshift operator notation

$$X_t = \psi(B)Z_t$$

where

$$\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j.$$

The process is termed non-anticipating or causal if $\psi_j = 0$ for all $j < 0$, so that

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}.$$

2.4 Linear Processes

Note that for the linear process to be sensible we need the sum

$$\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

to converge in some appropriate fashion; in this context we require mean-square convergence, denoted

$$X_t \overset{\text{m.s.}}{=} \sum_{j=0}^{\infty} \psi_j Z_{t-j}.$$

2.4 Linear Processes

We require for the definition

$$\lim_{n \longrightarrow \infty} \mathbb{E}\left[\left(X_t - \sum_{j=-n}^{n} \psi_j Z_{t-j}\right)^2\right] = 0$$

For finite $n \geqslant 0$,

$$\left(\sum_{j=-n}^{n} \psi_j Z_{t-j}\right)^2 = \sum_{j=-n}^{n} \psi_j^2 Z_{t-j}^2 + \sum_{j=-n}^{n} \sum_{k \neq j} \psi_j \psi_k Z_{t-j} Z_{t-k}$$

2.4 Linear Processes

Taking expectations, as $\{Z_t\}$ is a white noise process so that $\mathbb{E}[Z_{t-j}Z_{t-k}] = 0$ if $j \neq k$

$$\mathbb{E}\left[\left(\sum_{j=-n}^{n} \psi_j Z_{t-j}\right)^2\right] = \sigma_Z^2 \sum_{j=-n}^{n} \psi_j^2.$$

2.4 Linear Processes

Now

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty \quad \Longrightarrow \quad \sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$$

so therefore the partial sums (with limits $\pm n$) are also finite for all n.

If

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

then we term the sequence $\{\psi_j\}$ absolutely summable.

2.4 Linear Processes

Note also that by the Cauchy-Schwarz inequality

$$\mathbb{E}[|Z_t|] \leqslant \sqrt{\mathbb{E}[Z_t^2]} = \sigma,$$

therefore

$$\mathbb{E}[|X_t|] \leqslant \sum_{j=-\infty}^{\infty} |\psi_j| \mathbb{E}[|Z_{t-j}|] \leqslant \sigma \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

so that

$$\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

is convergent (almost surely).

2.4 Linear Processes

The operator $\psi(B)$ is termed a linear filter; it acts on the stationary (white noise) process $\{Z_t\}$ to produce the stationary (autocorrelated) process $\{X_t\}$.

The operator can also be applied to any stationary process.

2.4 Linear Processes

Suppose that $\{Y_t\}$ is a weakly stationary, zero mean process with ACVF γ_Y, and $\psi(B)$ is the above linear filter with absolutely summable coefficients. Then the process

$$X_t = \psi(B)Y_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j}$$

is also weakly stationary with mean zero. The ACVF of $\{X_t\}$ is

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h + j - k).$$

2.4 Linear Processes

If $\{Y_t\}$ is a white noise process, so that $\{X_t\}$ is a linear process, then

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \sigma_Y^2.$$

To see this, note that

$$\mathbb{E}[X_t] \leqslant \mathbb{E}[|X_t|] \leqslant \sum_{j=-\infty}^{\infty} |\psi_j| |\mathbb{E}[Y_t]| < \infty$$

so that $\mathbb{E}[X_t] = 0$ for all t. Then

$$\mathbb{E}[X_t X_{t+h}] = \mathbb{E}\left[\left(\sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} \right) \left(\sum_{k=-\infty}^{\infty} \psi_k Y_{t+h-k} \right) \right]$$

Multiplying out yields

$$\begin{aligned}\mathbb{E}[X_t X_{t+h}] &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \mathbb{E}\left[Y_{t-j} Y_{t+h-k}\right] \\ &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(j+h-k).\end{aligned}$$

If $\{Y_t\}$ is white noise, then $\gamma_Y(h) = \sigma_Y^2$ if $h = 0$ and zero otherwise, in which case

$$\gamma_X(h) = \mathbb{E}[X_t X_{t+h}] = \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \sigma_Y^2.$$

2.4 Linear Processes

If $\alpha(B)$ and $\beta(B)$ are two linear filters, each of which has absolutely summable coefficients, and

$$\psi(B) = \alpha(B)\beta(B) = \beta(B)\alpha(B)$$

then $\psi(B)$ is also a linear filter, with

$$\psi_j = \sum_{k=-\infty}^{\infty} \alpha_k \beta_{j-k} = \sum_{k=-\infty}^{\infty} \beta_k \alpha_{j-k}$$

and if $\{Y_t\}$ is a stationary process, then

$$X_t = \psi(B)Y_t = \alpha(B)\beta(B)Y_t$$

defines a stationary process $\{X_t\}$. Note that $W_t = \beta(B)Y_t$ is also stationary, and $X_t = \alpha(B)W_t$.

Example: AR(1) process

Let $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, $|\phi| < 1$, and

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

with $\mathbb{E}[Z_s X_t] = 0$ for $s > t$. That is, $\{X_t\}$ is the solution of the equation

$$X_t - \phi X_{t-1} = Z_t$$

or

$$(1 - \phi B) X_t = Z_t.$$

Example: AR(1) process

Consider the linear process

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

Note that

$$\sum_{j=0}^{\infty} |\phi|^j = \frac{1}{1-|\phi|} < \infty$$

hence the conditions of absolute summability hold.

Example: AR(1) process

This is a solution to the equation above as

$$
\begin{aligned}
X_t &= \phi X_{t-1} + Z_t \\
&= \phi\left(\phi X_{t-2} + Z_{t-1}\right) + Z_t && \text{recursion} \\
&\vdots \\
&= \lim_{n \longrightarrow \infty} \left\{ \phi^n X_{t-n} + \sum_{j=0}^{n-1} \phi^j Z_{t-j} \right\} \\
&= \sum_{j=0}^{\infty} \phi^j Z_{t-j}
\end{aligned}
$$

Example: AR(1) process

Also, for $h \geqslant 0$,

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \sigma_Z^2 = \sum_{j=0}^{\infty} \phi^j \phi^{j+h} \sigma_Z^2 = \sigma_Z^2 \sum_{j=0}^{\infty} \phi^{2j+h}$$

that is

$$\gamma_X(h) = \frac{\sigma_Z^2 \phi^h}{1 - \phi^2}$$

Example: AR(1) process

If $\{Y_t\}$ is another stationary solution to the AR(1) equation, then, for any $k \geqslant 0$

$$\begin{aligned}
\mathbb{E}\left[\left(Y_t - \sum_{j=0}^{k} \phi^j Z_{t-j}\right)^2\right] &= \mathbb{E}\left[\left(\sum_{j=k+1}^{\infty} \phi^j Z_{t-j}\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{l=0}^{\infty} \phi^{l+k+1} Z_{t-k-1-l}\right)^2\right] \\
&= \phi^{2(k+1)} \mathbb{E}\left[\left(\sum_{l=0}^{\infty} \phi^{l} Z_{t-k-1-l}\right)^2\right] \\
&= \phi^{2(k+1)} \mathbb{E}\left[\left(Y_{t-k-1}\right)^2\right]
\end{aligned}$$

Example: AR(1) process

The right hand side of this equation converges to zero as $k \longrightarrow \infty$, as $|\phi| < 1$, and $\mathbb{E}\left[Y_{t-k-1}^2\right] < \infty$.

Therefore Y_t and X_t are equal in mean square

$$Y_t \overset{\text{m.s.}}{=} X_t$$

and therefore $\{X_t\}$ is essentially the unique stationary solution to the AR(1) equation

$$X_t - \phi X_{t-1} = Z_t$$

with $|\phi| < 1$ and $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$.

Example: AR(1) process

Therefore for any stationary AR process there is an essentially unique MA(∞) representation, with $\psi_j = \phi^j$ for $j \geqslant 0$, and $\psi_j = 0$ for $j < 0$.

Example: Non-stationary AR(1) process

If $|\phi| = 1$, say

$$X_t = X_{t-1} + Z_t \qquad \text{or} \qquad X_t = -X_{t-1} + Z_t$$

then X_t is not stationary. Clearly

$$\text{Var}[X_t] = \text{Var}[\pm X_{t-1}] + \text{Var}[Z_t] > \text{Var}[X_{t-1}]$$

for all t.

Non-causal AR(1): In the AR(1) formulation, suppose $|\phi| > 1$, that is

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

so that

$$\frac{1}{\phi} X_t = X_{t-1} + \frac{1}{\phi} Z_t$$

or

$$X_{t-1} = \phi_* X_t + Z_t^*$$

where $\{Z_t^*\} \sim \text{WN}(0, \sigma_Z^2/\phi^2)$ is defined by

$$Z_t^* = -\frac{1}{\phi} Z_t$$

and $\phi_* = 1/\phi$, so that $|\phi_*| < 1$.

2.4 Linear Processes

Therefore, by considering a reverse time formulation, we can construct a stationary solution to the AR(1) equation. It takes the linear process form

$$X_t = \sum_{j=0}^{\infty} \phi_*^j Z_{t+1+j}^* = -\sum_{j=0}^{\infty} \phi^{-(j+1)} Z_{t+1+j}$$

that is, X_t is defined in terms of future values of $\{Z_t\}$.

The ACVF can be computed by the usual methods. Such a process is termed **non-causal**.

2.4 Linear Processes

Note: we have for the AR(1) process that

$$(1 - \phi B)X_t = Z_t \qquad \therefore \qquad X_t = (1 - \phi B)^{-1} Z_t$$

and by a geometric series expansion,

$$X_t = \left\{ \sum_{j=0}^{\infty} \phi^j B^j \right\} Z_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$$

as before.

2.5 Autoregressive Processes

Suppose that $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and that the equation

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$$

is used to define (implicitly) $\{X_t\}$. This definition does not ensure (or even suggest) that the solution can be found, or is unique or stationary.

This process is termed the autoregressive process of order p (AR(p)). We have that

$$\mathbb{E}[X_t] = \mathbb{E}[Z_t] = 0.$$

Example: AR(2) process

If $p = 2$ we have

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = Z_t \qquad \text{or} \qquad \Phi(B) X_t = Z_t.$$

where $\Phi(B) = (1 - \phi_1 B - \phi_2 B^2)$, and the following linear process representation obtained by noting that

$$X_t = (1 - \phi_1 B - \phi_2 B^2)^{-1} Z_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

where ψ_j is the coefficient of B^j in the series expansion of

$$(1 - \phi_1 B - \phi_2 B^2)^{-1}$$

Example: AR(2) process

Consider the factorization

$$\Phi(z) = (1 - \phi_1 z - \phi_2 z^2) = (1 - \xi_1 z)(1 - \xi_2 z).$$

where z is an arbitrary complex number, and (ξ_1, ξ_2) are a complex conjugate pair. In general,

$$(1 - \xi z)^{-1}$$

has radius of convergence $|\xi| < 1$ when $|z| = 1$.

2.5 Autoregressive Processes

Example: AR(2) process

(ξ_1, ξ_2) are the reciprocal roots (that is, the reciprocals of the roots) of the equation

$$\Phi(z) = 0$$

then the series expansion of

$$\{(1 - \xi_1 z)(1 - \xi_2 z)\}^{-1}$$

is convergent for $|z| = 1$ if and only if

$$|\xi_1| = |\xi_2| < 1.$$

2.5 Autoregressive Processes

Example: AR(2) process

We have that

$$\Phi(B)X_t = Z_t \qquad \therefore \qquad X_t = \{\Phi(B)\}^{-1} Z_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

where ψ_j is the coefficient of z^j in the series expansion of

$$\{\Phi(z)\}^{-1} = \{(1-\xi_1 z)(1-\xi_2 z)\}^{-1} = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \xi_1^j \xi_2^k z^{j+k}$$

that is

$$\psi_j = \sum_{k=0}^{j} \xi_1^k \xi_2^{j-k}$$

Example: AR(2) process

Note that

$$|\psi_j| \leqslant \sum_{k=0}^{j} |\xi_1|^k |\xi_2|^{j-k} = (j+1)M^j$$

where $M = \max\{|\xi_1|, |\xi_2|\} < 1$. Therefore

$$\sum_{j=0}^{\infty} |\psi_j| \leqslant \sum_{j=0}^{\infty} (j+1)M^j = \frac{1}{(1-M)^2} < \infty.$$

Hence $\{X_t\}$ is a linear process with absolutely summable coefficients.

2.5 Autoregressive Processes

The extension to the AR(p) case is straightforward: write

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = \prod_{j=1}^{p} (1 - \xi_j z)$$

and by an identical argument to above, to ensure stationarity, we require that the reciprocal roots

$$\xi_1, \xi_2, \ldots, \xi_p$$

all satisfy $|\xi_j| < 1$.

Equivalently, if we let $\eta_j = \xi_j^{-1}$ denote the roots of $\Phi(z) = 0$, then we have

$$\Phi(z) = \prod_{j=1}^{p} (\eta_j - z)$$

and require $|\eta_j| > 1$ for each j.

2.5 Autoregressive Processes

If $|\xi_j| > 1$ for at least one j, then no causal stationary solution to the AR(p) equation exists, but a similar approach to the AR(1) case allows us to define a (unique) non-causal stationary solution.

Hence the only case that leads to non-stationary solutions is when $|\xi_j| = 1$ for at least one j. In the simplest case $\xi_j = 1$, which corresponds to the factor $(1 - B)$ in the AR polynomial; however any complex root with modulus 1 yields a non-stationary solution, that is

$$\xi_j = e^{i\omega}$$

for $-\pi < \omega \leqslant \pi$, where $i = \sqrt{-1}$.

Example: AR(2) process

Suppose

$$X_t - \frac{13}{4}X_{t-1} + \frac{3}{4}X_{t-2} = Z_t.$$

Then

$$\Phi(z) = 1 - \frac{13}{4}z + \frac{3}{4}z^2 = (1 - 3z)(1 - z/4)$$

Therefore the reciprocal roots are 3 and $1/4$. Therefore no causal stationary solution exists.

Example: AR(2) process

Suppose

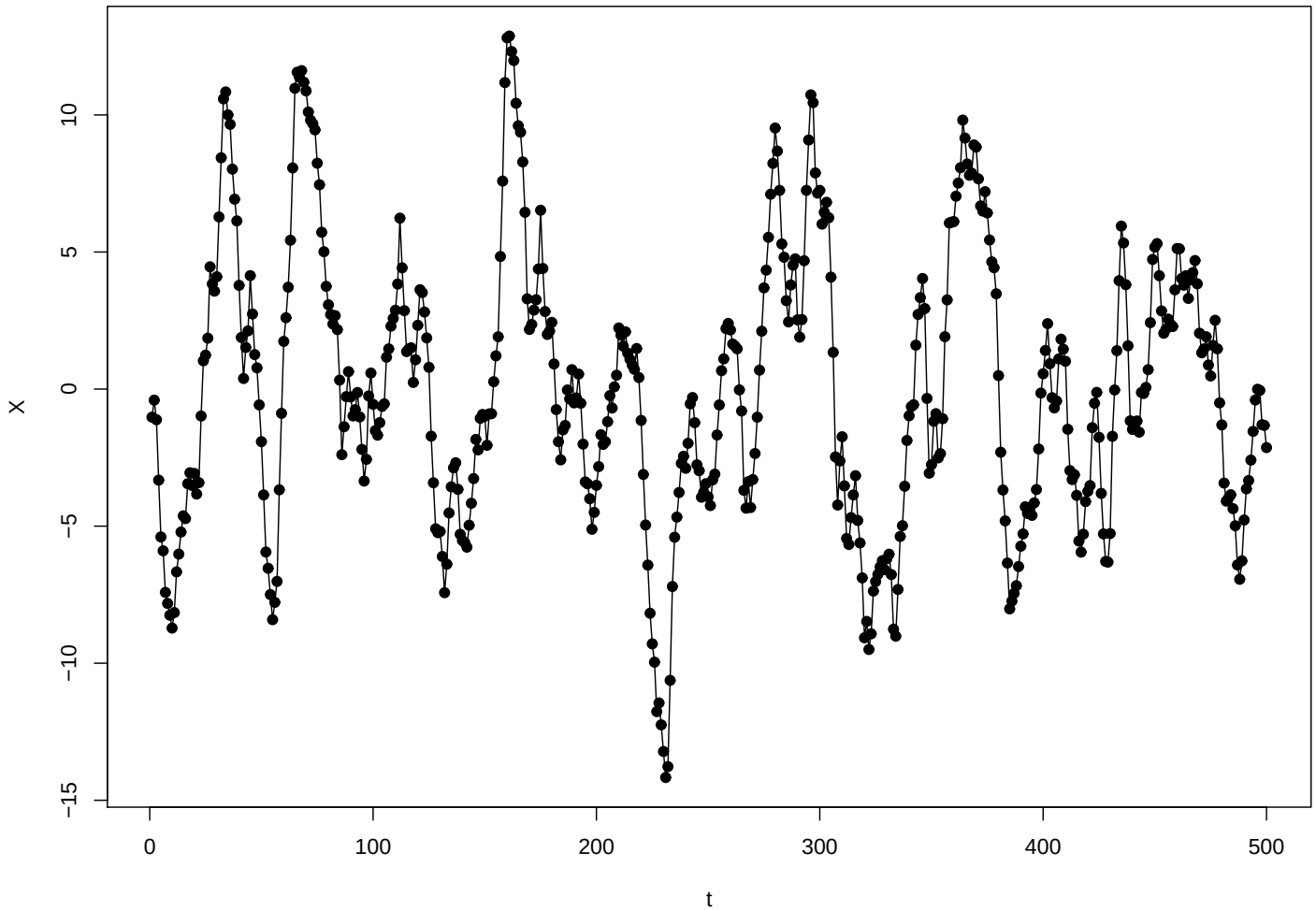
$$X_t - \frac{31}{20}X_{t-1} + \frac{3}{5}X_{t-2} = Z_t.$$

Then

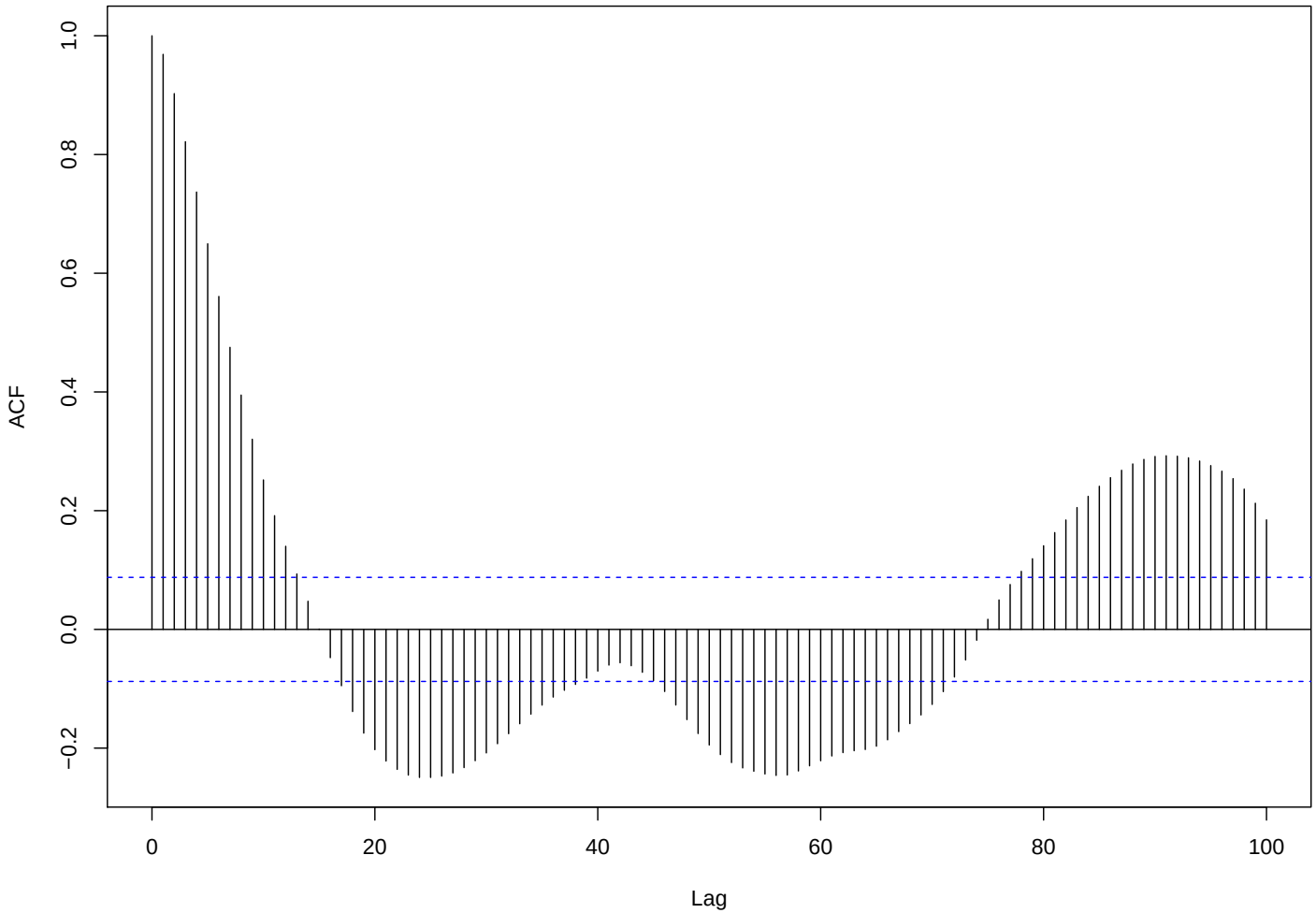
$$\Phi(z) = 1 - \frac{31}{20}z + \frac{3}{5}z^2 = (1 - 4z/5)(1 - 3z/4)$$

Therefore the reciprocal roots are $4/5$ and $3/4$. Therefore a causal stationary solution does exist.

2.5 Autoregressive Processes



2.5 Autoregressive Processes



2.5 Autoregressive Processes

Code: we can use the functions in `R` to perform key calculations

- `polyroot` to compute the roots of $\Phi(z) = 0$.
- `arima.sim` to simulate a time series.
- `ARMAtoMA` to compute the ψ_j values in the MA(∞) linear process formulation
- `ARMAacf` to compute the theoretical ACF.

In these functions, the representation

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$$

is used, so that the coefficients in the above model are $\phi_1 = 31/20$ and $\phi_2 = -3/5$.

2.5 Autoregressive Processes

```r
polyroot(c(3/5,-31/20,1))  #Finding the roots (coefs in increasing term order)

## [1] 0.75+2.041443e-14i 0.80-2.041443e-14i

#Numerical solutions for roots are in complex number form, but imaginary part is negligible.

#Using the roots of a quadratic formula
aval<-1
bval<--31/20
cval<-3/5
root.vals<-(-bval+c(-1,1)*sqrt(bval^2-4*aval*cval))/(2*aval)
root.vals

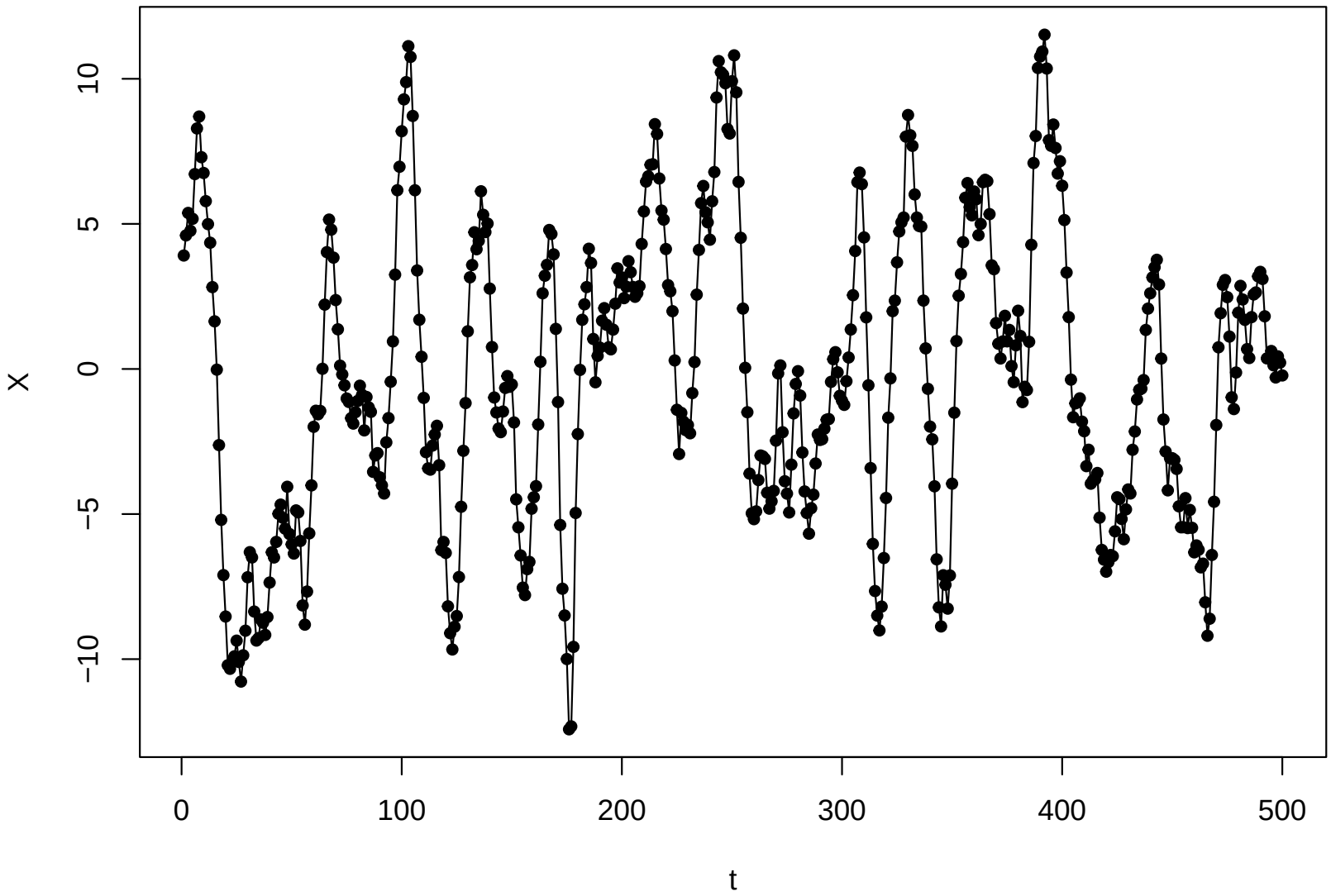
## [1] 0.75 0.80

1/root.vals

## [1] 1.333333 1.250000

set.seed(32)
X<-arima.sim(n=500,model=list(ar=c(31/20,-3/5)))
par(mar=c(4,4,2,2))
plot(1:500,X,pch=19,cex=0.8,xlab='t')
lines(1:500,X)
```

2.5 Autoregressive Processes



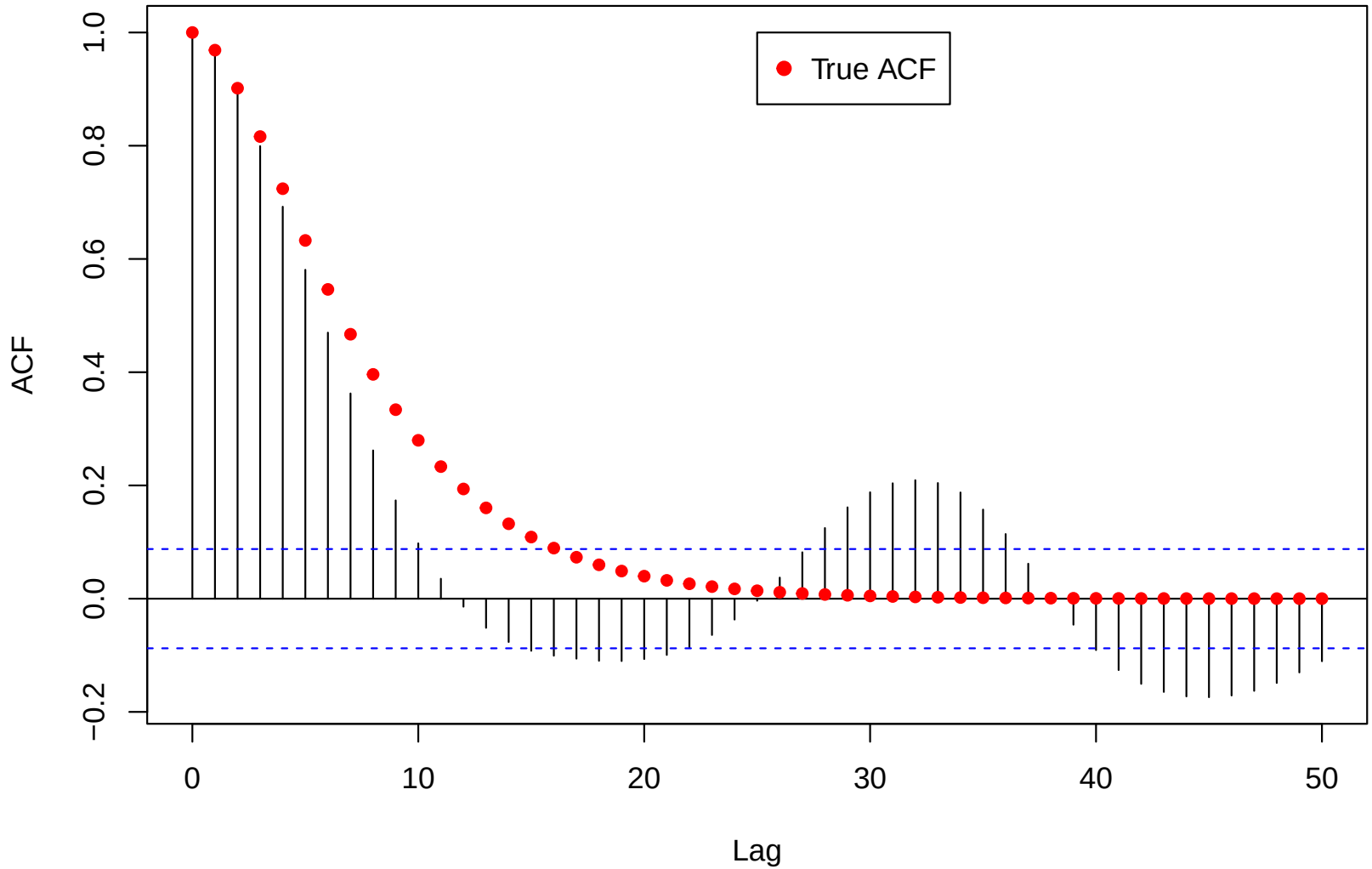
2.5 Autoregressive Processes

```r
ARMAtoMA(ar=c(31/20,-3/5),lag.max=20)  #Linear Process representation (first 20 coefs)

##  [1] 1.5500000 1.8025000 1.8638750 1.8075063 1.6833097 1.5246263 1.3531849
##  [8] 1.1826608 1.0212134 0.8732842 0.7408625 0.6243663 0.5232503 0.4364182
## [15] 0.3624980 0.3000210 0.2475338 0.2036647 0.1671601 0.1368993

par(mar=c(4,4,1,2))
true.acf<-ARMAacf(ar=c(31/20,-3/5),lag.max=50)
acf(X,main='ACF of X',lag.max=50)
points(0:50,true.acf,col='red',pch=19,cex=0.8)
legend(25,1,c('True ACF'),pch=19,col='red')
```

2.5 Autoregressive Processes

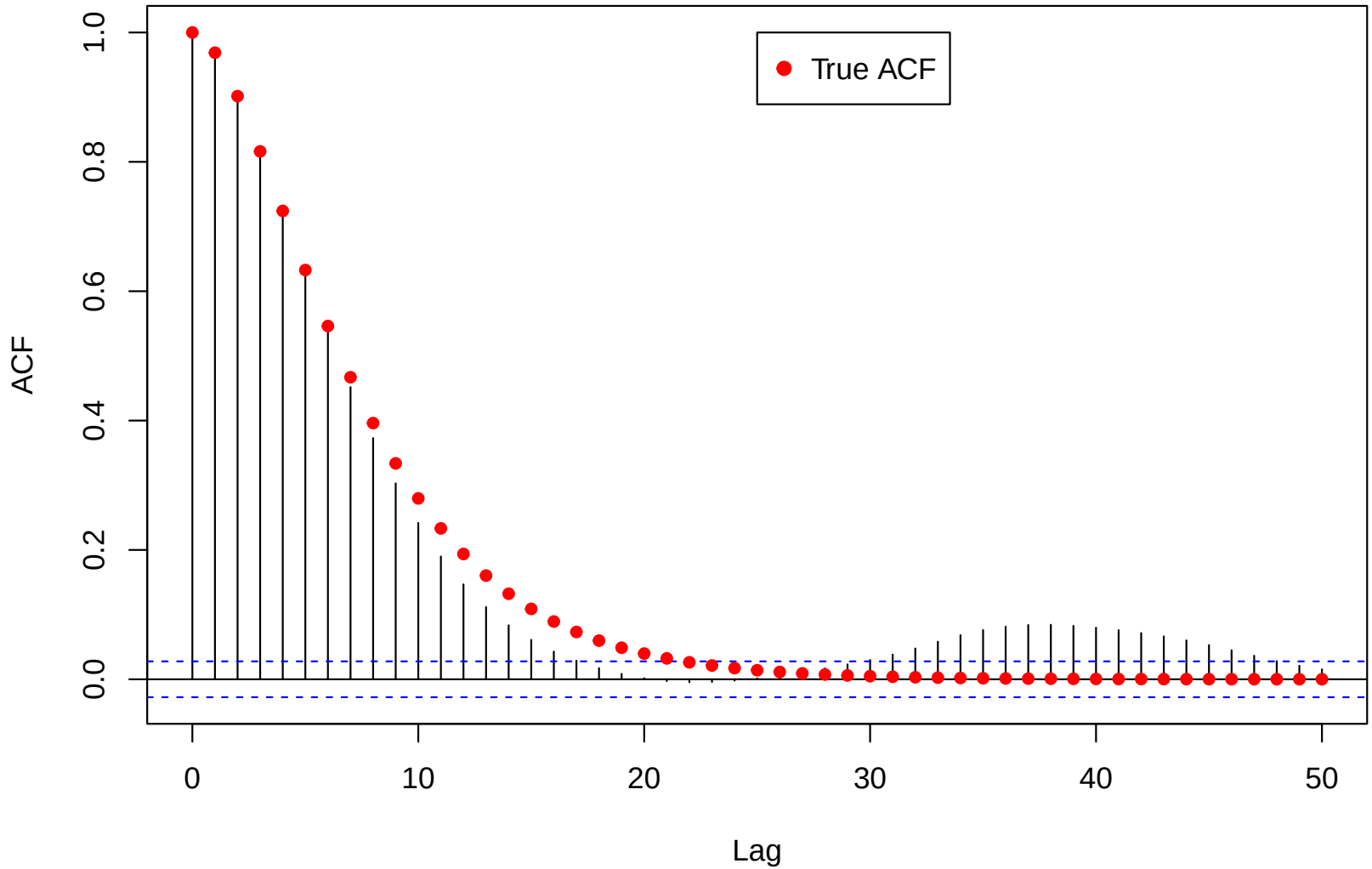


2.5 Autoregressive Processes

Larger sample size: $n = 5000$.

```
set.seed(32)
X<-arima.sim(n=5000,model=list(ar=c(31/20,-3/5)))
par(mar=c(4,4,1,2))
true.acf<-ARMAacf(ar=c(31/20,-3/5),lag.max=50)
acf(X,main='ACF of X',lag.max=50)
points(0:50,true.acf,col='red',pch=19,cex=0.8)
legend(25,1,c('True ACF'),pch=19,col='red')
```

2.5 Autoregressive Processes



2.6 ARMA Processes

If $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, we have

- **Moving Average:**

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

  - stationary
  - q-dependent; $\gamma_X(h) = 0$ for $h > q$

2.6 ARMA Processes

- **Autoregression:**

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$$

  - stationary provided the roots of the AR polynomial

$$\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = \prod_{j=1}^{p} (1 - \xi_j z)$$

    satisfy $|\xi_j| \neq 1$ for each j; causal if $|\xi_j| < 1$ for each j.
  - yields an MA(∞) representation.

2.6 ARMA Processes

$\{X_t\}$ is an ARMA(p, q) process if it is stationary and satisfies

$$\Phi(B)X_t = \Theta(B)Z_t$$

where $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and

$$\begin{aligned}
\Phi(z) &= 1 - \phi_1 z - \cdots - \phi_p z^p = \prod_{j=1}^{p}(1 - \xi_j z) \\
\Theta(z) &= 1 + \theta_1 z + \cdots + \theta_q z^q = \prod_{j=1}^{q}(1 - \omega_j z)
\end{aligned}$$

Example: ARMA$(1,1)$

Suppose $\{X_t\}$ satisfies

$$(1-\phi B)X_t = (1+\theta B)Z_t$$

This is the ARMA$(1,1)$ process. Suppose that $|\phi| < 1$, and let

$$\alpha(B) = \{\Phi(B)\}^{-1} = \sum_{j=0}^{\infty} \phi^j B^j$$

Example: ARMA$(1,1)$

Then

$$X_t = \alpha(B)\Theta(B)Z_t = \Psi(B)Z_t$$

where

$$\Psi(z) = \alpha(z)\Theta(z) = (1 + \phi z + \phi^2 z^2 + \cdots)(1 + \theta z)$$

Example: ARMA$(1,1)$

Therefore

$$\begin{aligned} \psi_0 &= 1 \\ \psi_j &= (\theta+\phi)\phi^{j-1},\ j \geqslant 1. \end{aligned}$$

so

$$X_t = Z_t + (\theta+\phi)\sum_{j=1}^{\infty}\phi^{j-1}Z_{t-j}$$

Note: if $|\phi| > 1$, can construct a stationary non-causal version.

Invertibility: If

$$\Phi(B)X_t = \Theta(B)Z_t$$

then we may also write

$$\{\Theta(B)\}^{-1}\Phi(B)X_t = Z_t$$

provided the expansion

$$\beta(B) = \{\Theta(B)\}^{-1} = \sum_{j=0}^{\infty} \beta_j B^j$$

converges.

From above, we have

$$\Theta(z) = \prod_{j=1}^{q} (1 - \omega_j z)$$

so therefore expansion exists provided $|\omega_j| < 1$ converges. Such a process is termed **invertible.**

Example: ARMA$(1,1)$

In this case

$$\Theta(B) = 1 + \theta B \qquad \beta(B) = 1 - \theta B + \theta^2 B^2 - \cdots$$

and we require $|\theta| < 1$. Let $\pi(B) = \beta(B)\Phi(B)$, where

$$\pi_0 = 1$$

$$\pi_j = -(\theta + \phi)(-\theta)^{j-1},\ j \geqslant 1.$$

We then have an AR(∞) representation

$$\pi(B)X_t = Z_t.$$

2.7 ARMA processes: Estimation

Estimation for ARMA processes, and indeed all stationary processes, can be achieved in a non-parametric fashion using moment-based estimation.

For example, we may estimate expectation $\mu = \mathbb{E}[X_t]$ using estimator

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_t = \overline{X}$$

2.7 ARMA processes: Estimation

Recall that we require an estimator to have good statistical properties. In the non IID case, properties of this estimator are more difficult to study; it is clear that, under stationarity,

$$\mathbb{E}[\overline{X}] = \mu$$

so we have an unbiased estimator, but the variance of this estimator must be computed with care.

2.7 ARMA processes: Estimation

$$\begin{aligned}
\mathrm{Var}[\overline{X}] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}[X_i, X_j] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_X(|i-j|) \\
&= \frac{1}{n^2}\sum_{h=-n}^{n}(n-|h|)\gamma_X(h) \\
&= \frac{1}{n}\sum_{h=-n}^{n}\left(1-\frac{|h|}{n}\right)\gamma_X(h)
\end{aligned}$$

As $n \longrightarrow \infty$, this variance should stay finite, otherwise the performance of the estimator will be poor.

2.7 ARMA processes: Estimation

A necessary condition is that

$$\gamma_X(h) \longrightarrow 0 \qquad \text{as } h \longrightarrow \infty,$$

but this is not sufficient, for example if $\gamma_X(h) \sim 1/h$ as $h \longrightarrow \infty$, the sum diverges.

2.7 ARMA processes: Estimation

If $\text{Var}[\overline{X}] \longrightarrow 0$, then $\overline{X} \longrightarrow \mu$ in mean square. If

$$\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty$$

the asymptotic variance is the probability limit of $n\text{Var}[\overline{X}]$,

$$n\text{Var}[\overline{X}] \xrightarrow{p} \sum_{h=-\infty}^{\infty} \gamma_X(h)$$

Note that it is not necessarily the case that

$$\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty.$$

2.7 ARMA processes: Estimation

For a Gaussian series $\{X_t\}$

$$\sqrt{n}(\overline{X} - \mu) \sim \mathcal{N}\left(0, \sum_{h=-n}^{n} \left(1 - \frac{|h|}{n}\right) \gamma_X(h)\right)$$

whereas for non-Gaussian stationary series this result holds as an asymptotic approximation, provided the autocovariance series is absolutely summable.

2.7 ARMA processes: Estimation

In practice, $\gamma_X(h)$ must be estimated from the data, and in the asymptotic case, the infinite sum must be truncated;

$$\sum_{h=-l}^{l} \left(1 - \frac{|h|}{n}\right) \widehat{\gamma}(h)$$

commonly this is done at $l = \sqrt{n}$.

2.7 ARMA processes: Estimation

Example: AR(1) process

Suppose that $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and

$$X_t - \mu = \phi(X_{t-1} - \mu) + Z_t$$

with $|\phi| < 1$. Then

$$\gamma_X(h) = \frac{\sigma_Z^2 \phi^{|h|}}{1 - \phi^2}$$

Example: AR(1) process

Then

$$\sum_{h=-\infty}^{\infty} |\gamma_X(h)| = \sum_{h=-\infty}^{\infty} \gamma_X(h)$$

$$= \frac{\sigma_Z^2}{1-\phi^2}\left[1 + 2\sum_{h=1}^{\infty}\phi^h\right]$$

$$= \frac{\sigma_Z^2}{(1-\phi)^2}$$

2.7 ARMA processes: Estimation

Estimation of the autocovariance function $\gamma_X(h)$ can also be achieved by moment-based methods

$$\widehat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+h} - \overline{x})(x_t - \overline{x}).$$

This is a biased estimator of $\gamma_X(h)$, but the bias is low (order $o(1/n)$).

Note that the $k \times k$ symmetrix matrix $\widehat{\mathbf{\Gamma}}_k$ with

$$[\widehat{\mathbf{\Gamma}}_k]_{ij} = \widehat{\gamma}(|i-j|)$$

is non-negative definite for each $k \geqslant 1$.

2.7 ARMA processes: Estimation

The autocorrelation function $\rho_X(h)$ can be estimated by

$$\widehat{\mathbf{R}}_k = \frac{1}{\widehat{\gamma}(0)}\widehat{\mathbf{\Gamma}}_k$$

The properties of this estimator can be established, and can be used to construct hypothesis tests for each h that $\rho_X(h)$ is zero based on a z-score approach.

2.7 ARMA processes: Estimation

In particular, under the null hypothesis that the process is a white-noise process, we have approximately that

$$\sqrt{n}[\mathbf{R}_k]_{1,h} = \sqrt{n}\hat{\rho}(h) \sim \mathcal{N}(0,1) \qquad h \geqslant 1$$

Global (or portmanteau) tests, that assess multiple h can be constructed based on a Chi-squared distribution approximation. See for example the Box-Ljung test (`Box.test` in R).

2.7 ARMA processes: Estimation

Note: The matrices $\mathbf{\Gamma}_k, \widehat{\mathbf{\Gamma}}_k$ and other related matrices are symmetric Toeplitz matrices (that is, they are constant along diagonals).

$$\Gamma_k = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \gamma(3) & \cdots & \gamma(k-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(k-2) \\ \gamma(2) & \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(k-3) \\ \gamma(3) & \gamma(2) & \gamma(1) & \gamma(0) & \cdots & \gamma(k-4) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(k-1) & \gamma(k-2) & \gamma(k-3) & \gamma(k-4) & \cdots & \gamma(0) \end{bmatrix}$$

2.7 ARMA processes: Estimation

Note: The matrices $\boldsymbol{\Gamma}_k, \widehat{\boldsymbol{\Gamma}}_k$ and other related matrices are symmetric Toeplitz matrices (that is, they are constant along diagonals).

$$\Gamma_k = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \gamma(3) & \cdots & \gamma(k-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(k-2) \\ \gamma(2) & \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(k-3) \\ \gamma(3) & \gamma(2) & \gamma(1) & \gamma(0) & \cdots & \gamma(k-4) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(k-1) & \gamma(k-2) & \gamma(k-3) & \gamma(k-4) & \cdots & \gamma(0) \end{bmatrix}$$

2.8 Forecasting for Stationary Processes

Forecasting for stationary processes utilizes the autocovariance structure of the process to optimize predictions. Suppose we wish to forecast X_{n+h} given the observed data

$$x_{1:n} = (x_1, \ldots, x_n).$$

We focus on linear predictors, that is, linear combinations of past values.

We will focus on model-free procedures, and utilize moment properties rather distributional properties of $\{X_t\}$.

2.8 Forecasting for Stationary Processes

We utilize the minimum Mean-Square Error (MSE) criterion, and choose constants $a_{0:n} = (a_0, a_1, \ldots, a_n)^\top$ to minimize

$$\mathbb{E}\left[\left(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n-i+1}\right)^2\right]$$

where the expectation is over the joint distribution of all the random quantities.

We denote the linear predictor using the n-data prediction operator notation

$$P_n X_{n+h} = a_0 + \sum_{i=1}^{n} a_i X_{n-i+1}$$

2.8 Forecasting for Stationary Processes

By differentiation, we need to solve simultaneously

$$\mathbb{E}\left[\left(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n-i+1}\right)\right] = 0$$

and, for $j = 1, \ldots, n$

$$\mathbb{E}\left[\left(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n-i+1}\right) X_{n-j+1}\right] = 0$$

From the first equation, we see that we must choose

$$a_0 = \mu\left(1 - \sum_{i=1}^{n} a_i\right)$$

where $\mu = \mathbb{E}[X_t]$.

2.8 Forecasting for Stationary Processes

In vector notation, the resulting system of n simultaneous equations is written

$$\mathbf{\Gamma}_n a = \gamma_{h:(h+n-1)} \tag{1}$$

where

$$\begin{aligned} \mathbf{\Gamma}_n &= [\gamma_X(i-j)]_{ij} \qquad (n \times n) \\ a &= (a_1, \ldots, a_n)^\top \qquad (n \times 1) \\ \gamma_{h:(h+n-1)} &= (\gamma_X(h), \ldots, \gamma_X(h+n-1))^\top \qquad (n \times 1) \end{aligned}$$

with $a_0 \equiv a_0(a_1, \ldots, a_n)$ given by the previous expression. This is a linear system in $(a_1, \ldots, a_n)$; in general the solution will also depend on h.

2.8 Forecasting for Stationary Processes

Denote the solution to (1) by

$$a(n,h) = (a(1,h),\ldots,a(n,h))^\top$$

so that

$$P_n X_{n+h} = \mu + \sum_{i=1}^{n} a(i,h)(X_{n-i+1} - \mu).$$

Note that, by construction

$$\begin{aligned}
\mathbb{E}[(X_{n+h} - P_n X_{n+h})] &= 0 \\
\mathbb{E}[(X_{n+h} - P_n X_{n+h})X_j] &= 0 \qquad j = 1,\ldots,n
\end{aligned}$$

The minimum value of the MSE achieved by $a(n,h)$ is

$$\mathbb{E}[(X_{n+h} - P_n X_{n+h})^2] = \gamma_X(0) - a(n,h)^\top \boldsymbol{\gamma}_{h:(h+n-1)}$$

2.8 Forecasting for Stationary Processes

The formula

$$\mathbf{\Gamma}_n a = \gamma_{h:(h+n-1)}$$

depends only on the autocovariance function. Direct solution of this equation is possible for most practical situations

$$a(n,h) = \mathbf{\Gamma}_n^{-1}\gamma_{h:(h+n-1)}$$

can be computed using `solve` in R. However, the form of $\mathbf{\Gamma}_n$ (symmetric, Toeplitz) can make it relatively easy to invert.

2.8 Forecasting for Stationary Processes

Example: AR(1)

Recall that

$$[\mathbf{\Gamma}_n]_{ij} = \phi^{|i-j|}$$

so that

$$\boldsymbol{\gamma}_{h:(h+n-1)} = (\phi^h, \phi^{h+1}, \ldots, \phi^{h+n-1})^\top$$

Then the matrix inversion can be carried out analytically, yielding

$$a(n,h) = (\phi^h, 0, 0, \ldots, 0)^\top$$

which is an intuitively reasonable result. It follows that

$$P_n X_{n+1} = \phi X_n \qquad\qquad P_n X_{n+2} = \phi^2 X_n$$

and so on.

2.8 Forecasting for Stationary Processes

Let $Y, W_1, \ldots, W_n$ be random variables with

$$\mathbb{E}[Y] = \mu_Y \quad \mathbb{E}[W_i] = \mu_i$$

and

$$\mathrm{Var}[Y], \mathrm{Var}[W_i], \mathrm{Cov}[Y, W_i], \mathrm{Cov}[W_i, W_j] < \infty.$$

Write $W_{n:1} = (W_n, \ldots, W_1)^\top$, and

$$\boldsymbol{\gamma} = \mathrm{Cov}(Y, W_{n:1}) \qquad (n \times 1)$$
$$\boldsymbol{\Gamma} = \mathrm{Cov}(W_{n:1}, W_{n:1}) \qquad (n \times n)$$

2.8 Forecasting for Stationary Processes

Denote the best linear predictor of Y, given $W_1, \ldots, W_n$ by

$$\widehat{P}(Y|W) = \mu_Y + a^\top (W_{n:1} - \mu_{n:1})$$

where a solves $\mathbf{\Gamma} a = \boldsymbol{\gamma}$. The minimum MSE is then

$$\mathbb{E}[(Y - \widehat{P}(Y|W))^2] = \mathrm{Var}[Y] - a^\top \gamma.$$

2.8 Forecasting for Stationary Processes

Let U and V be random variables with $\mathbb{E}[U^2], \mathbb{E}[V^2] < \infty$, and $\beta, \alpha_1, \ldots, \alpha_n$ be constants. Then

1. $\widehat{P}(U|W_{n:1}) = \mathbb{E}[U] + a^\top (W_{n:1} - \mu_{n:1})$.
2. $\mathbb{E}[U - \widehat{P}(U|W_{n:1})] = \mathbb{E}[(U - \widehat{P}(U|W_{n:1}))W_{n:1}] = \mathbf{0}$.
3. $\mathbb{E}[(U - \widehat{P}(U|W_{n:1}))^2] = \mathrm{Var}[U] - a^\top \gamma$.
4. $\widehat{P}$ is a linear operator

$$\widehat{P}(\alpha_1 U + \alpha_2 V + \beta|W_{n:1}) = \alpha_1 \widehat{P}(U|W_{n:1}) + \alpha_2 \widehat{P}(V|W_{n:1}) + \beta.$$

2.8 Forecasting for Stationary Processes

5. $\widehat{P}(\alpha^\top W_{n:1} + \beta | W_{n:1}) = \alpha^\top W_{n:1} + \beta$.
6. If $\gamma = 0$, $\widehat{P}(U|W_{n:1}) = \mathbb{E}[U]$.
7. Coherent iterated prediction: If $W = (W_{n:1}, W^*_{n:1})$, then

$$\widehat{P}(U|W) = \widehat{P}(\widehat{P}(U|W_{n:1}, W^*_{n:1})|W_{n:1}).$$

2.8 Forecasting for Stationary Processes

Example: AR(p) process

If $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, and

$$\Phi(B)X_t = Z_t$$

such that $\{X_t\}$ is stationary and causal. Then for each n

$$\begin{aligned} P_{n-1}X_n &= P_n[\phi_1 X_{n-1} + \cdots + \phi_p X_{n-p} + Z_n] \\ &= \phi_1 X_{n-1} + \cdots + \phi_p X_{n-p} \end{aligned}$$

by 4., 5. and 6.

2.9 Levinson-Durbin Algorithm

For a stationary process $\{X_t\}$, suppose we write

$$P_n X_{n+1} = \sum_{j=1}^{n} \phi_{n,j} X_{n-j+1} = \Phi_n^\top X_{1:n}$$

say as the general form of a linear predictor, where

$$\Phi_n = (\phi_{n,1}, \ldots, \phi_{n,n})^\top \qquad (n \times 1)$$

is the coefficient vector resulting from n data.

2.9 Levinson-Durbin Algorithm

By the arguments above, the optimal MSE choice for ϕ_n is

$$\phi_n = \mathbf{\Gamma}_n^{-1}\boldsymbol{\gamma}_n$$

where

$$[\mathbf{\Gamma}_n]_{ij} = \gamma_X(|i-j|) \qquad \boldsymbol{\gamma}_n = (\gamma_X(1), \ldots, \gamma_X(n))^\top.$$

Let the minimum MSE be denoted

$$v_n = \mathbb{E}[(X_{n+1} - P_n X_{n+1})^2] = \gamma_X(0) - \phi_n^\top \boldsymbol{\gamma}_n.$$

2.9 Levinson-Durbin Algorithm

We seek a recursion for ϕ_n; suppose ϕ_{n-1} and v_{n-1} have been computed, and define ϕ_n by setting the nth component as

$$\phi_{n,n} = \frac{1}{v_{n-1}} \left[\gamma_X(n) - \Phi_{n-1}^\top \gamma_{(n-1):1} \right]$$

and the first $n-1$ components via

$$\Phi_{n,1:(n-1)} = \Phi_{n-1} - \phi_{n,n} \Phi_{n-1}^{\mathrm{R}}$$

where

$$\Phi_{n-1}^{\mathrm{R}} = (\phi_{n-1,n-1}, \ldots, \phi_{n-1,1})^\top$$

is the reversed version of ϕ_{n-1}. Finally, set

$$v_n = v_{n-1}(1 - \phi_{n,n}^2)$$

2.9 Levinson-Durbin Algorithm

To initialize the recursion, set

$$\phi_{1,1} = \rho_X(1) = \frac{\gamma_X(1)}{\gamma_X(0)} \qquad v_0 = \gamma_X(0) \qquad v_1 = \gamma_X(0)(1 - \phi_{1,1}^2)$$

To see that the recursion produces a valid solution, first set $\mathbf{R}_n$ to be the autocorrelation matrix $\mathbf{R}_n = \mathbf{\Gamma}_n / \gamma_X(0)$, and define

$$\boldsymbol{\rho}_n = \frac{1}{\gamma_X(0)} \boldsymbol{\gamma}_n = (\rho_X(1), \ldots, \rho_X(n))^\top.$$

We verify the inductive step

$$\mathbf{R}_n \boldsymbol{\phi}_n = \boldsymbol{\rho}_n \qquad \Longrightarrow \qquad \mathbf{R}_{n+1} \boldsymbol{\phi}_{n+1} = \boldsymbol{\rho}_{n+1}$$

2.9 Levinson-Durbin Algorithm

Clearly the relation holds for $n = 1$, as

$$\mathbf{R}_1 = 1 \qquad \boldsymbol{\phi}_1 = \phi_{1,1} = \rho_X(1) \qquad \boldsymbol{\rho}_1 = \rho_X(1).$$

We assume the relation holds for $n = k$, and verify that it holds for $n = k + 1$. Now

$$\mathbf{R}_{k+1} = \begin{bmatrix} \mathbf{R}_k & \boldsymbol{\rho}_k^{\mathrm{R}} \\ \{\boldsymbol{\rho}_k^{\mathrm{R}}\}^\top & 1 \end{bmatrix}$$

where

$$\boldsymbol{\rho}_k^{\mathrm{R}} = (\rho_X(k), \ldots, \rho_X(1))^\top = \frac{1}{\gamma_X(0)} \boldsymbol{\gamma}_k^{\mathrm{R}}$$

2.9 Levinson-Durbin Algorithm

Recall that

$$\mathbf{R}_{k+1} = \frac{1}{\gamma_X(0)} \begin{bmatrix} \gamma_X(0) & \gamma_X(1) & \gamma_X(2) & \cdots & \gamma_X(k) \\ \gamma_X(1) & \gamma_X(0) & \gamma_X(1) & \cdots & \gamma_X(k-1) \\ \gamma_X(2) & \gamma_X(1) & \gamma_X(0) & \cdots & \gamma_X(k-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_X(k) & \gamma_X(k-1) & \gamma_X(k-2) & \cdots & \gamma_X(0) \end{bmatrix}$$

and note that

$$\mathbf{R}_k \phi_k = \boldsymbol{\rho}_k \qquad \Longrightarrow \qquad \mathbf{R}_k \phi_k^{\mathrm{R}} = \boldsymbol{\rho}_k^{\mathrm{R}}.$$

$$
\begin{aligned}
\mathbf{R}_{k+1}\boldsymbol{\phi}_{k+1} &= \begin{bmatrix} \mathbf{R}_k & \boldsymbol{\rho}_k^{\mathrm{R}} \\ \{\boldsymbol{\rho}_k^{\mathrm{R}}\}^\top & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}_k - \phi_{k+1,k+1}\boldsymbol{\phi}_k^{\mathrm{R}} \\ \phi_{k+1,k+1} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{R}_k\boldsymbol{\phi}_k - \phi_{k+1,k+1}\mathbf{R}_k\boldsymbol{\phi}_k^{\mathrm{R}} + \phi_{k+1,k+1}\boldsymbol{\rho}_k^{\mathrm{R}} \\ \{\boldsymbol{\rho}_k^{\mathrm{R}}\}^\top\boldsymbol{\phi}_k - \phi_{k+1,k+1}\{\boldsymbol{\rho}_k^{\mathrm{R}}\}^\top\boldsymbol{\phi}_k^{\mathrm{R}} + \phi_{k+1,k+1} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{R}_k\boldsymbol{\phi}_k \\ \{\boldsymbol{\rho}_k^{\mathrm{R}}\}^\top\boldsymbol{\phi}_k + \phi_{k+1,k+1}\big(1 - \{\boldsymbol{\rho}_k^{\mathrm{R}}\}^\top\boldsymbol{\phi}_k^{\mathrm{R}}\big) \end{bmatrix}
\end{aligned}
$$

as $\mathbf{R}_k\boldsymbol{\phi}_k^{\mathrm{R}} = \boldsymbol{\rho}_k^{\mathrm{R}}$.

2.9 Levinson-Durbin Algorithm

Now by definition

$$\phi_{k+1,k+1} = \frac{1}{v_k}\left[\gamma_X(k+1) - \phi_k^\top \gamma_k^{\mathrm{R}}\right]$$

and by definition

$$\begin{aligned} v_k = \gamma_X(0) - \phi_k^\top \gamma_k &= \gamma_X(0) - \{\phi_k^{\mathrm{R}}\}^\top \gamma_k^{\mathrm{R}} \\ &= \gamma_X(0)(1 - \{\boldsymbol{\rho}_k^{\mathrm{R}}\}^\top \phi_k^{\mathrm{R}}) \end{aligned}$$

so therefore

$$\frac{1}{v_k}(1 - \{\boldsymbol{\rho}_k^{\mathrm{R}}\}^\top \phi_k^{\mathrm{R}}) = \frac{1}{\gamma_X(0)}$$

Therefore

$$\phi_{k+1,k+1} = \frac{1}{\gamma_X(0)}\left[\gamma_X(k+1) - \boldsymbol{\phi}_k^\top \boldsymbol{\gamma}_k^{\mathrm{R}}\right] = \rho_X(k+1) - \boldsymbol{\phi}_k^\top \boldsymbol{\rho}_k^{\mathrm{R}}$$

and hence

$$\begin{aligned}
\mathbf{R}_{k+1}\boldsymbol{\phi}_{k+1} &= \begin{bmatrix} \mathbf{R}_k\boldsymbol{\phi}_k \\ \{\boldsymbol{\rho}_k^{\mathrm{R}}\}^\top \boldsymbol{\phi}_k + \rho_X(k+1) - \boldsymbol{\phi}_k^\top \boldsymbol{\rho}_k^{\mathrm{R}} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\rho}_k \\ \rho_X(k+1) \end{bmatrix} \\
&= \boldsymbol{\rho}_{k+1}
\end{aligned}$$

and the recursion holds.

2.9 Levinson-Durbin Algorithm

For the prediction MSE recursion: using the recursion for ϕ_n,

$$\begin{aligned}
v_n &= \mathbb{E}[(X_{n+1} - \phi_n^\top X_{n:1})^2] = \gamma_X(0) - \phi_n^\top \gamma_n \\
&= \gamma_X(0) - \phi_{n-1}^\top \gamma_{n-1} + \phi_{n,n} \left\{\phi_{n-1}^{\mathrm{R}}\right\}^\top \gamma_{n-1} - \phi_{n,n}\gamma_X(n) \\
&= v_{n-1} + \phi_{n,n} \left(\left\{\phi_{n-1}^{\mathrm{R}}\right\}^\top \gamma_{n-1} - \gamma_X(n) \right) \\
&= v_{n-1} - \phi_{n,n}^2 (\gamma_X(0) - \phi_{n-1}^\top \gamma_{n-1}) \\
&= v_{n-1}(1 - \phi_{n,n}^2).
\end{aligned}$$

where line 2 follows from line 1 by substituting in the formula for ϕ_n given by the recursion.

2.10 Partial Autocorrelation Function (PACF)

Define the function $\alpha_X(h)$ by $\alpha_X(0) = 1$ and

$$\alpha_X(h) = \phi_{h,h} \qquad h = 1, 2, \ldots, n$$

where $\phi_{h,h}$ is the value returned as the coefficient on the hth step of the Levinson-Durbin recursion.

This function is termed the partial autocorrelation function, or PACF, which is uniquely determined by the ACF γ_X.

2.10 Partial Autocorrelation Function (PACF)

Note also that as v_n is non-negative, and

$$v_n = v_{n-1}(1 - \phi_{n,n}^2)$$

we must have that $\phi_{n,n}^2 \leqslant 1$ for all n, and hence that

$$v_n \leqslant v_{n-1}$$

so the optimal prediction MSEs form a decreasing sequence.

2.10 Partial Autocorrelation Function (PACF)

For random variables X, Y, Z, the partial correlation of X and Y, given Z is defined by

$$\text{Corr}[(X - \mathbb{E}[X|Z]), (Y - \mathbb{E}[Y|Z])]$$

that is, the correlation between the residuals from a regression of X on Z with those from a regression of Y on Z.

2.10 Partial Autocorrelation Function (PACF)

For process $\{X_t\}$, if we let $X \equiv X_t$, $Y = X_{t+h}$, and $Z = X_{(t+1):(t+h-1)}$.

$$\begin{aligned}\alpha_X(h) &= \phi_{h,h} \\ &= \text{Corr}\big[\big(X_t - \widehat{P}(X_t|X_{(t+1):(t+h-1)})\big), \\ &\qquad \big(X_{t+h} - \widehat{P}(X_{t+h}|X_{(t+1):(t+h-1)})\big)\big]\end{aligned}$$

The PACF computes the correlation between prediction residuals for X_t and X_{t+h}, using $X_{(t+1):(t+h-1)}$ to make the predictions.

2.10 Partial Autocorrelation Function (PACF)

Example: AR(p)

Suppose $\{X_t\} \sim \text{AR}(p)$, with parameters $\phi_1, \ldots, \phi_p$, and consider $h > p$. Then

$$\widehat{P}(X_{t+h}|X_{(t+1):(t+h-1)}) = \phi_1 X_{t+h-1} + \cdots + \phi_p X_{t+h-p}$$

and

$$\widehat{P}(X_t|X_{(t+1):(t+h-1)}) = \widehat{X}_t = \phi_1 X_{t+1} + \cdots + \phi_p X_{t+p}$$

so therefore

$$\begin{aligned} X_{t+h} - \widehat{P}(X_{t+h}|X_{(t+1):(t+h-1)}) &= Z_{t+h} \\ X_t - \widehat{P}(X_t|X_{(t+1):(t+h-1)}) &= Z_t \end{aligned}$$

and as $\text{Corr}[Z_t, Z_{t+h}] = 0$, the PACF is zero for $h > p$.

2.10 Partial Autocorrelation Function (PACF)

The Levinson-Durbin algorithm provides a method for solving the linear system

$$\mathbf{R}_n \phi_n = \boldsymbol{\rho}_n$$

by recursion, for arbitrary n, and given ACF $\gamma_X(h)$ without the need for large matrix inversion. It exploits the symmetric Toeplitz nature of an autocovariance matrix, and is an order n^2 algorithm.

Other algorithms for inverting this type of matrix are typically higher order; the Cholesky decomposition

$$\mathbf{R}_n = \mathbf{L}_n \mathbf{L}_n^\top$$

with $\mathbf{L}_n$ lower triangular is order n^3. It does not exploit the Toeplitz nature of $\mathbf{R}_n$.

2.11 The Innovations Algorithm

Another approach to forecasting proves feasible even for the non-stationary case. Suppose that $\{X_t\}$ is zero mean process with $\mathbb{E}[|X_t|^2]$, and $\mathbb{E}[X_i X_j] = \kappa(i,j)$. Denote

$$\widehat{X}_n = \begin{cases} 0 & n = 1 \\ P_{n-1} X_n & n = 2, 3, \ldots. \end{cases}$$

and let

$$U_n = X_n - \widehat{X}_n$$

denote the one step prediction error. Then

$$U_{1:n} = \mathbf{A}_n X_{1:n}$$

where $\mathbf{A}_n$ contains the coefficients $a_{i,j}$ that appear in the optimal linear predictor.

2.11 The Innovations Algorithm

$$\mathbf{A}_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{1,1} & 1 & 0 & \cdots & 0 \\ a_{2,2} & a_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1,n-1} & a_{n-1,n-2} & a_{n-1,n-3} & \cdots & 1 \end{bmatrix}$$

This lower triangular matrix is non-singular and let

$$\mathbf{A}_n^{-1} = \mathbf{C}_n$$

where $\mathbf{C}_n$ is also lower triangular.

2.11 The Innovations Algorithm

Let

$$\mathbf{C}_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 1 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 1 \end{bmatrix}$$

Then, recalling that $X_{1:n} = \mathbf{C}_n U_{1:n}$,

$$X_{1:n} = \mathbf{C}_n U_{1:n} = \mathbf{C}_n (X_{1:n} - \widehat{X}_{1:n}).$$

Also

$$\widehat{X}_{1:n} = X_{1:n} - U_{1:n} = \mathbf{C}_n U_{1:n} - U_{1:n} = \mathbf{\Theta}_n (X_{1:n} - \widehat{X}_{1:n})$$

2.11 The Innovations Algorithm

$$\mathbf{\Theta}_n = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 0 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 0 \end{bmatrix}$$

Therefore

$$\widehat{X}_{n+1} = \begin{cases} 0 & n = 0 \\ \sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \widehat{X}_{n+1-j}) & n = 1, 2, \ldots. \end{cases}$$

2.11 The Innovations Algorithm

Therefore the one-step predictions

$$\widehat{X}_1, \widehat{X}_2, \ldots, \widehat{X}_n, \ldots$$

can be computed as weighted sums of previous prediction errors

$$X_1 - \widehat{X}_1, X_2 - \widehat{X}_2, \ldots, X_n - \widehat{X}_n, \ldots$$

provided the entries in the matrix $\mathbf{\Theta}_n$ are known. Note that by the previous results, the sequence $\{D_n\}$ defined by

$$D_n = X_n - \widehat{X}_n$$

is an **uncorrelated** sequence.

Another recursive approach avoids the need to carry out matrix inversion.

2.11 The Innovations Algorithm

Initialize $v_0 = \kappa(1,1) = \mathbb{E}[X_1^2]$, then for each n set

$$\theta_{n,n-k} = \frac{1}{v_k}\left[\kappa(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right]$$

for $0 \leqslant k < n$, and

$$v_n = \kappa(n+1,n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j$$

That is, we compute

$$\begin{array}{cccc} v_0 & & & \\ \theta_{1,1} & v_1 & & \\ \theta_{2,2} & \theta_{2,1} & v_2 & \\ \theta_{3,3} & \theta_{3,2} & \theta_{3,1} & v_3 \end{array}$$

and so on.

2.11 The Innovations Algorithm

For h-step ahead prediction

$$P_n X_{n+h} = \sum_{j=h}^{n+h-1} \theta_{n+j-1,j}\left(X_{n+h-j} - \hat{X}_{n+j-j}\right)$$

and

$$v_{n+h} = \kappa(n+h, n+h) - \sum_{j=h}^{n+h-1} \theta^2_{n+j-1,j} v_{n+h-j-1}$$

In this case, the innovations algorithm must be run **forward** $h-1$ steps.

2.11 The Innovations Algorithm

Note: It is possible to extend the idea of optimal linear prediction from the finite case to the infinite, that is, to find constants $\{\beta_j\}$ such that

$$\widetilde{P}X_{n+h} = \sum_{j=1}^{\infty} \beta_j X_{n+1-j}.$$

The infinite system of moment constraints

$$\mathbb{E}[(X_{n+h} - \widetilde{P}X_{n+h})X_{n+1-i}] = 0 \qquad i = 1, 2, \ldots$$

yield the infinite system of equations

$$\sum_{j=1}^{\infty} \gamma_X(i-j)\beta_j = \gamma_X(h+i-1) \qquad i = 1, 2,$$

2.12 The Wold Decomposition

Any process $\{X_t\}$ where

$$X_n - \widetilde{P}_{n-1} X_n = 0$$

in mean square, that is

$$\mathbb{E}\left[(X_n - \widetilde{P}_{n-1} X_n)^2\right] = 0$$

for all n is termed deterministic

2.12 The Wold Decomposition

Example: Simple Deterministic Process

Suppose U, V are zero mean and uncorrelated variables with variance σ^2, and let

$$X_t = U\cos(\omega t) + V\sin(\omega t)$$

for some $\omega \in (0, \pi)$. Then for each integer n,

$$X_n = 2\cos(\omega)X_{n-1} - X_{n-2} = \widetilde{P}_{n-1}X_n$$

say, and hence

$$X_n - \widetilde{P}_{n-1}X_n = 0$$

2.12 The Wold Decomposition

The Wold Decomposition gives a fundamental representation of stationary processes.

If $\{X_t\}$ is stationary and non-deterministic, then $\{X_t\}$ has the representation

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t \qquad \text{(Wold)}$$

where

- $\{Z_t\} \sim \text{WN}(0, \sigma^2)$
- $\{V_t\}$ is deterministic
- $\{Z_t\}$ and $\{V_t\}$ are uncorrelated
- $\psi_0 = 1$, $\sum_{j=1}^{\infty} \psi_j^2 < \infty$.

2.12 The Wold Decomposition

We also have that

- $Z_t = \widetilde{P}_t Z_t$ for all t
- $V_t = \widetilde{P}_s V_t$ for all s, t.

and that the representation in (Wold) is unique. Furthermore

- for each t, $Z_t = X_t - \widetilde{P}_{t-1} X_t$;
- for each j, ψ_j satisfies

$$\psi_j = \frac{\mathbb{E}[X_t Z_{t-j}]}{\mathbb{E}[Z_t^2]};$$

- for each t, $V_t = X_t - \sum_{j=0}^{\infty} \psi_j Z_{t-j}$.

Chapter 3
ARMA Models

3. ARMA Models

We study further properties of ARMA(p, q) processes.

Consider the ARMA(p, q) process $\{X_t\}$,

$$\Phi(B)X_t = \Theta(B)Z_t$$

where $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$,

3.1 Basic Properties

Here

$$\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = \prod_{j=1}^{p} (1 - \xi_j z)$$

$$\Theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q = \prod_{j=1}^{q} (1 - \omega_j z)$$

We assume that there are no common factors in $\Phi(z)$ in $\Theta(z)$.

- $\{X_t\}$ is causal and stationary if and only if $|\xi_j| < 1, j = 1, \ldots, p$
- $\{X_t\}$ is invertible if and only if $|\omega_j| < 1, j = 1, \ldots, q$.

3.1 Basic Properties

Note that by using the power series expansion of the generating function

$$\Psi(z) = \frac{\Theta(z)}{\Phi(z)}$$

we can construct the linear process (MA(∞)) representation

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

We must have

$$(1 - \phi_1 z - \cdots - \phi_p z^p)(\psi_0 + \psi_1 z + \cdots) = (1 + \theta_1 z + \cdots + \theta_q z^q)$$

so equating coefficients, we have $\psi_0 = 1$, and for $1, 2, \ldots$, we have the recursion

$$\psi_j = \theta_j + \sum_{k=1}^{p} \phi_k \psi_{j-k}$$

By a similar argument, under invertibility, we can construct an AR(∞) representation

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

where $\pi_0 = 1$, and $j = 0, 1, 2, \ldots$

$$\pi_j = -\phi_j - \sum_{k=1}^{q} \theta_k \pi_{j-k}$$

3.1 Basic Properties

ACF: The ACF $\gamma_X(h)$ is given by previous results for linear processes; for integer h

$$\gamma_X(h) = \mathbb{E}[X_t X_{t+h}] = \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$$

Example: MA(q)

By the linear process representation, $\psi_0 = 1$, $\psi_j = \theta_j$ for $j = 1, \ldots, q$, and $\psi_j = 0$ for $j > q$, so therefore

$$\gamma_X(h) = \sigma_Z^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} \qquad 0 \leqslant h \leqslant q$$

and zero otherwise.

Example: ARMA$(1,1)$

Consider the ARMA$(1,1)$ process based on a WN$(0,\sigma_Z^2)$ process, with $|\phi| < 1$. Then recall that

$$\begin{aligned}\psi_0 &= 1\\ \psi_j &= (\theta+\phi)\phi^{j-1},\ j \geqslant 1.\end{aligned}$$

gives the MA(∞) representation. Note that for $j \geqslant 2$,

$$\psi_j = \phi\psi_{j-1}.$$

Example: ARMA$(1,1)$ stationary and causal

$h = 0$:

$$\gamma_X(0) = \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j^2$$

$$= \sigma_Z^2 \left[\psi_0^2 + \sum_{j=1}^{\infty} \psi_j^2 \right]$$

$$= \sigma_Z^2 \left[1 + (\theta + \phi)^2 \sum_{j=0}^{\infty} \phi^{2j} \right]$$

$$= \sigma_Z^2 \left[1 + \frac{(\theta + \phi)^2}{1 - \phi^2} \right]$$

Example: ARMA$(1,1)$ stationary and causal

$h = 1$:

$$\gamma_X(1) = \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+1}$$

$$= \sigma_Z^2 \left[(\theta + \phi) + (\theta + \phi)^2 \sum_{j=1}^{\infty} \phi^{2j-1} \right]$$

$$= \sigma_Z^2 \left[(\theta + \phi) + (\theta + \phi)^2 \phi \sum_{j=0}^{\infty} \phi^{2j} \right]$$

$$= \sigma_Z^2 \left[(\theta + \phi) + \frac{\phi(\theta + \phi)^2}{1 - \phi^2} \right]$$

Example: ARMA$(1,1)$ stationary and causal

For $h \geqslant 2$

$$\gamma_X(h) = \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}$$

$$= \sigma_Z^2 \phi \sum_{j=0}^{\infty} \psi_j \psi_{j+h-1} \qquad \psi_{j+h} = \phi \psi_{j+h-1}$$

$$= \phi \gamma_X(h-1)$$

3.1 Basic Properties

For a direct computation for the general ARMA case, note that

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

Multiplying through by X_{t-h}, $h = 0, 1, 2, \ldots$:

$$\begin{aligned} \text{LHS} \quad & X_t X_{t-h} - \phi_1 X_{t-1} X_{t-h} - \cdots - \phi_p X_{t-p} X_{t-h} \\ \text{RHS} \quad & Z_t X_{t-h} + \theta_1 Z_{t-1} X_{t-h} + \cdots + \theta_q Z_{t-q} X_{t-h}. \end{aligned}$$

3.1 Basic Properties

Taking expectations we have for $j = 0, 1, \ldots, p$

$$\mathbb{E}[X_{t-j}X_{t-h}] = \mathbb{E}[X_{t+h-j}X_t] = \gamma_X(h-j)$$

so for the left-hand side we have

$$\gamma_X(h) - \phi_1\gamma_X(h-1) - \cdots - \phi_p\gamma_X(h-p).$$

For the right hand side, using the MA(∞) form

$$X_{t-h} = \sum_{l=0}^{\infty} \psi_l Z_{t-h-l}.$$

Thus for $j = 0, 1, \ldots, q$,

$$\mathbb{E}[X_{t-h} Z_{t-j}] = \mathbb{E}\left[\sum_{l=0}^{\infty} \psi_l Z_{t-h-l} Z_{t-j}\right] = \sum_{l=0}^{\infty} \psi_l \mathbb{E}\left[Z_{t-h-l} Z_{t-j}\right]$$

Now

$$\mathbb{E}\left[Z_{t-h-l} Z_{t-j}\right] = \begin{cases} \sigma_Z^2 & l = j - h \\ 0 & \text{otherwise} \end{cases}$$

Therefore, for fixed h, as $\psi_j = 0$ if $j < 0$, we have

$$\mathbb{E}[X_{t-h} Z_{t-j}] = \sigma_Z^2 \psi_{j-h}$$

for $j = h, \ldots, q$, and zero otherwise. Hence, writing $\theta_0 = 1$, we have

$$\mathbb{E}\left[\sum_{j=0}^{q} \theta_j Z_{t-j} X_{t-h}\right] = \sigma_Z^2 \sum_{j=h}^{q} \theta_j \psi_{j-h} = \sigma_Z^2 \sum_{j=0}^{q-h} \theta_{j+h} \psi_j$$

Note that if $h > q$, $\theta_{j+h} = 0$ for all $j \geqslant 0$, so in this case the right hand side is zero.

3.1 Basic Properties

If $m = \max\{p, q+1\}$, we therefore have the infinite system of equations

$$\gamma_X(h) - \sum_{j=1}^{p} \phi_j \gamma_X(h-j) = \begin{cases} \sigma_Z^2 \sum_{j=0}^{m} \theta_{j+h}\psi_j & 0 \leqslant h < m \\ 0 & h \geqslant m \end{cases}$$

This system in the unknowns

$$\gamma_X(0), \gamma_X(1), \gamma_X(2), \ldots$$

for fixed parameters

$$\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$$

can be solved using two approaches.

3.1 Basic Properties

- **Direct approach:** the equations form a system known as linear difference equations which has an analytic solution, expressed via the roots $\zeta_1, \ldots, \zeta_p$ of the AR polynomial equation $\Phi(z) = 0$

$$\gamma_X(h) = \sum_{j=1}^{p} \alpha_j \zeta_j^{-h} \qquad h \geqslant m - p$$

where the initial conditions determine $\alpha_1, \ldots, \alpha_p$ and

$$\gamma_X(0), \ldots, \gamma_X(h - p - 1).$$

- **Recursive approach:** The first $p+1$ equations form a linear system in the autocovariances $\gamma_X(0), \ldots, \gamma_X(p)$, which can be solved using matrix inversion, Levinson-Durbin, or innovations algorithm.

  The remaining autocovariances $\gamma_X(p+1), \gamma_X(p+2), \ldots$ can be computed using the recursion.

3.1 Basic Properties

In R, the function ARMAacf computes the ACF for given parameter settings

```
ARMAacf(ar = numeric(0), ma = numeric(0), lag.max = r, pacf = FALSE)
```

The function ARMAtoMA computes the MA(∞) coefficients.

3.2 Autocovariance Generating Function

The autocovariance generating function (ACVGF) can also be used for computing the ACV sequences.

For a real sequence $\{a_n\}$, the generating function

$$G(z) = \sum_n a_n z^n \qquad z \in \mathbb{C}$$

is defined for $|z|$ in a neighbourhood of 1, that is

$$\frac{1}{r} < |z| < r$$

for $r > 1$.

3.2 Autocovariance Generating Function

Suppose that $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

Suppose that there exists an $r > 1$ such that

$$\sum_{j=-\infty}^{\infty} |\psi_j| z^j \qquad \frac{1}{r} < |z| < r$$

is convergent. Then define the ACVGF for $\{X_t\}$ by

$$G_X(z) = \sum_{h=-\infty}^{\infty} \gamma_X(h) z^h \qquad \frac{1}{r} < |z| < r$$

3.2 Autocovariance Generating Function

As polynomials are defined by their coefficients, there is a 1-1 correspondence between $\gamma_X(h)$ and $G_X(z)$.

- This means that $\gamma_X(h)$ is uniquely identified by $G_X(z)$

3.2 Autocovariance Generating Function

Recall that

$$\gamma_X(h) = \sigma_Z^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|h|}$$

Therefore

$$\begin{aligned} G_X(z) &= \sigma_Z^2 \sum_{h=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|h|} z^h \\ &= \sigma_Z^2 \left[\sum_{j=-\infty}^{\infty} \psi_j^2 + \sum_{h=1}^{\infty} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \left(z^h + z^{-h} \right) \right] \\ &= \sigma_Z^2 \left[\sum_{j=-\infty}^{\infty} \psi_j z^j \right] \left[\sum_{h=-\infty}^{\infty} \psi_h z^{-h} \right] \\ &= \sigma_Z^2 \Psi(z) \Psi(z^{-1}), \text{ say.} \end{aligned}$$

3.2 Autocovariance Generating Function

For the stationary ARMA(p, q) model

$$\Phi(B)X_t = \Theta(B)Z_t$$

where $\Phi(z) \neq 0$ for $|z| = 1$, we have

$$X_t = \{\Phi(B)\}^{-1}\Theta(B)Z_t = \Psi(B)Z_t$$

say, so

$$G_X(z) = \sigma_Z^2\Psi(z)\Psi(z^{-1}) = \sigma_Z^2\frac{\Theta(z)\Theta(z^{-1})}{\Phi(z)\Phi(z^{-1})}$$

which is well defined and finite as the denominator is non zero.

3.2 Autocovariance Generating Function

Example: WN$(0, \sigma_Z^2)$

Note that if $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$,

$$G_Z(z) = \sigma_Z^2$$

and that this is a unique characteristic of white noise processes.

In general, the ACVF and ACVGF are not unique: in general, different ARMA processes can have the same ACVF and ACVGF.

3.2 Autocovariance Generating Function

We have seen that the ARMA(p, q) process

$$\Phi(B)X_t = \Theta(B)Z_t$$

is

- causal if the AR reciprocal roots satisfy

$$|\xi_j| < 1 \qquad j = 1, 2, \ldots, p$$

- invertible if the MA reciprocal roots satisfy

$$|\omega_j| < 1 \qquad j = 1, 2, \ldots, q$$

3.2 Autocovariance Generating Function

If the process is non-causal or non-invertible, we can derive a causal and invertible process with the same autocovariance sequence.

The new process is defined with respect to a different white noise process.

3.2 Autocovariance Generating Function

Example: MA(1)

If $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and

$$X_t = Z_t - 2Z_{t-1} = (1 - 2B)Z_t$$

so that

$$\Theta(z) = 1 - 2z,$$

as $|2| > 1$, the process is non-invertible. Define

$$Z_t^* = (1 - B/2)^{-1}(1 - 2B)Z_t.$$

Then

$$(1 - B/2)Z_t^* = (1 - 2B)Z_t = X_t.$$

3.2 Autocovariance Generating Function

Example: MA(1)

That is

$$\Phi(B)Z_t^* = \Theta(B)Z_t$$

where

$$\Phi(z) = 1 - \frac{1}{2}z \qquad \Theta(z) = 1 - 2z.$$

Thus

$$Z_t^* = \Psi(B)Z_t$$

where

$$\Psi(z) = \frac{\Theta(z)}{\Phi(z)} = \frac{(1-2z)}{(1-z/2)}$$

3.2 Autocovariance Generating Function

Example: MA(1)

We have

$$Z_t^* = (1 - B/2)^{-1} X_t = \sum_{j=0}^{\infty} \left(\frac{1}{2}\right)^j X_{t-j}$$

and the ACVGF for $\{Z_t^*\}$ is

$$\begin{aligned} G_{Z^*}(z) &= \sigma_Z^2 \Psi(z) \Psi(z^{-1}) \\ &= \sigma_Z^2 \frac{(1-2z)(1-2z^{-1})}{(1-z/2)(1-1/(2z))} \\ &= 4\sigma_Z^2 \frac{(1-2z)(z-2)}{(2-z)(2z-1)} \\ &= 4\sigma_Z^2 \end{aligned}$$

Example: MA(1)

Therefore $\{Z_t^*\} \sim \text{WN}(0, 4\sigma_Z^2)$, and

$$X_t = \Phi(B)Z_t^* = Z_t^* - \frac{1}{2}Z_{t-1}^*$$

has an invertible representation in terms of $\{Z_t^*\}$.

3.2 Autocovariance Generating Function

Consider the arbitrary stationary ARMA(p, q) process

$$\Phi(B)X_t = \Theta(B)Z_t$$

where $\Phi(z) \neq 0$, $\Theta \neq 0$ when $|z| = 1$

$$\Phi(z) = \prod_{j=1}^{p}(1 - \xi_j z)$$

$$\Theta(z) = \prod_{j=1}^{q}(1 - \omega_j z)$$

and

$$G_X(z) = \sigma_Z^2 \Psi(z)\Psi(z^{-1}) = \sigma_Z^2 \frac{\Theta(z)\Theta(z^{-1})}{\Phi(z)\Phi(z^{-1})}.$$

3.2 Autocovariance Generating Function

Then there exist polynomials

$$\Phi^*(z) \qquad \Theta^*(z)$$

and a white noise process $\{Z_t^*\}$ such that

$$\Phi^*(B)X_t = \Theta^*(B)Z_t^*$$

is a causal and invertible representation.

3.2 Autocovariance Generating Function

Suppose that the reciprocal roots are ordered such that

$$\xi_{r+1}, \ldots, \xi_p \qquad \omega_{s+1}, \ldots, \omega_q$$

lie outside of the unit circle. Then let

$$\Phi^*(z) = \Phi(z) \prod_{j=r+1}^{p} \left(\frac{1 - \xi_j z}{1 - \xi_j^{-1} z} \right)^{-1}$$

$$\Theta^*(z) = \Theta(z) \prod_{j=s+1}^{q} \left(\frac{1 - \omega_j z}{1 - \omega_j^{-1} z} \right)^{-1}$$

and consider

$$Z_t^* = \{\Theta^*(B)\}^{-1} \Phi^*(B) X_t$$

3.2 Autocovariance Generating Function

After some algebra, it follows that

$$G_{Z^*}(z) = \sigma_Z^2 \left\{ \prod_{j=r+1}^{p} |\xi_j|^{-2} \right\} \left\{ \prod_{j=s+1}^{q} |\omega_j|^2 \right\}$$

and hence that $\{Z_t^*\}$ is a white noise process and

$$\Phi^*(B)X_t = \Theta^*(B)Z_t^*$$

By construction, this is a causal and invertible representation of $\{X_t\}$.

3.3 Partial Autocorrelation

The PACF, $\alpha_X(h)$, defined by

$$\alpha_X(0) = 1$$

$$\alpha_X(h) = \phi_{h,h}, \quad h \geqslant 1.$$

$\phi_{h,h}$ is the last component of

$$\boldsymbol{\phi}_h = \boldsymbol{\Gamma}_h^{-1} \boldsymbol{\gamma}_h$$

We can carry out testing for the estimated PACF sequence as for ACF sequence.

3.4 Forecasting

The innovations algorithm can be used for the ARMA(p, q) model. If $m = \max\{p, q\}$, the one-step ahead forecasts are

$$\hat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}) & 1 \leqslant n \leqslant m \\ \sum_{j=1}^{p} \phi_j X_{n+1-j} + \sum_{j=1}^{q} \theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}) & n > m \end{cases}$$

where the $\theta_{n,j}$ are the quantities computed via the innovations algorithm.

3.5 Estimation

Estimation of the parameters of ARMA models proceeds using one of a number of different approaches

- Yule-Walker approach
- Burg's algorithm
- Maximum likelihood

Previous approaches use moment-based approaches in a non-parametric fashion.

Yule-Walker for AR(p) models: If

$$\Phi(B)X_t = Z_t$$

then using the previous technique, we premultiply by X_{t-h} for $h = 0, \ldots, p$ and take expectations, to yield

$$\gamma_X(h) = \phi_1\gamma_X(h-1) + \ldots + \phi_p\gamma_X(h-p) \qquad h = 0, \ldots, (p-1)$$

or equivalently, if $\boldsymbol{\phi}_p = (\phi_1, \ldots, \phi_p)^\top$,

$$\boldsymbol{\Gamma}_p\boldsymbol{\phi}_p = \boldsymbol{\gamma}_p$$

and

$$\sigma_Z^2 = \gamma_X(0) - \boldsymbol{\phi}_p^\top\boldsymbol{\gamma}_p$$

3.5 Estimation

Replacing the theoretical quantities by sample quantities, the equations yield estimates of ϕ_p and σ_Z^2; using the sample correlations, we have

$$\widehat{\phi}_p = \widehat{\mathbf{R}}_p^{-1}\widehat{\boldsymbol{\rho}}_p$$

As $n \longrightarrow \infty$,

$$\sqrt{n}(\widehat{\phi}_p - \phi_p) \sim \mathcal{AN}(0, \sigma_Z^2 \boldsymbol{\Gamma}_p)$$

The estimated parameters yield a stationary and causal procedure.

Note that to construct a stationary and causal AR(p) process we may use the relationship to define ϕ_p from $\gamma_X(h)$.

3.5 Estimation

Note: In the AR(p) case, if the models

$$\mathrm{AR}(1), \mathrm{AR}(2), \ldots, \mathrm{AR}(p), \ldots$$

are fitted, then the sequence of estimates

- $\widehat{\phi}_{1,1}$ for AR(1),
- $\widehat{\phi}_{2,1}, \widehat{\phi}_{2,2}$ for AR(2),
- ...
- $\widehat{\phi}_{p,1}, \ldots, \widehat{\phi}_{p,p}$ for AR(p)

coincide with the Levinson-Durbin coefficients.

The constants $\widehat{\phi}_{h,h}$, $h = 0, 1, 2, \ldots$ yield the PACF $\alpha_X(h)$.

3.5 Estimation

If $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and $\{X_t\}$ is an ARMA(p, q) model

$$\Phi(B)X_t = \Theta(B)Z_t$$

with a causal linear process representation

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

the equivalent calculation yields, for $0 \leqslant h \leqslant p + q$,

$$\gamma_X(h) - \phi_1\gamma_X(h-1) - \cdots - \phi_p\gamma_X(h-p) = \sigma_Z^2 \sum_{j=h}^{q} \theta_j\psi_{j-h}.$$

These equations are non-linear in the parameters.

Burg's Algorithm for AR(p) models: This algorithm proceeds via direct computation of the PACF series.

For $0 \leqslant i < n$, define

$$U_i(t) = X_{n+1+i-t} - \widehat{X}^{\mathrm{F}}_{n+1+i-t}$$

$$V_i(t) = X_{n+1+i-t} - \widehat{X}^{\mathrm{B}}_{n+1+i-t}$$

where $\widehat{X}^{\mathrm{F}}_{n+1+i-t}$ and $\widehat{X}^{\mathrm{B}}_{n+1+i-t}$ are the forward and backward forecasts of $X_{n+1+i-t}$ based on the preceding/following i observations.

Then $U_0(t) = V_0(t) = X_{n+1-t}$, and

$$U_i(t) = U_{i-1}(t-1) - \phi_{i,i} V_{i-1}(t)$$

$$V_i(t) = V_{i-1}(t) - \phi_{i,i} U_{i-1}(t-1).$$

We then choose $\phi_{i,i}$ to minimize

$$\frac{1}{2(n-i)} \sum_{t=i+1}^{n} (u_i(t)^2 + v_i(t)^2)$$

that is, the sums of forward and backward prediction errors.

This minimization can be achieved analytically, with the computation performed recursively.

3.5 Estimation

Maximum Likelihood Estimation: In most cases, maximum likelihood (ML) estimation provides the most efficient approach, and usually produces estimators have good small sample properties.

However, outside of the Normal case, the likelihood function is not often tractable.

3.5 Estimation

We consider first Gaussian time series $\{X_t\}$, presumed to be zero mean, with autocovariance

$$\mathbb{E}[X_i X_j] = \kappa(i,j)$$

Let

$$\widehat{X}_{t+1} = \mathbb{E}[X_{t+1}|X_1,\ldots,X_t] = \widehat{P}_t X_{t+1} \qquad t = 1,2,\ldots$$

which is always available for the Gaussian case.

3.5 Estimation

If $X_{1:t} = (X_1, \ldots, X_t)^\top$, then

$$\widehat{X}_{t+1} = \boldsymbol{\gamma}_{t+1,t:1}^\top \boldsymbol{\Gamma}_t^{-1} X_{1:t}$$

where

$$\boldsymbol{\gamma}_{t+1,t:1} = \mathbb{E}[X_{t+1} X_{t:1}] \qquad (t \times 1)$$

$$\boldsymbol{\Gamma}_t = \mathbb{E}[X_{1:t} X_{1:t}^\top] \qquad (t \times t)$$

are derived from the properties of the general multivariate normal distribution.

3.5 Estimation

Let

$$\mathbf{\Gamma}_n = \mathbb{E}[X_{1:n} X_{1:n}^\top]$$

then the likelihood for the parameters of $\mathbf{\Gamma}_n$ is

$$\mathcal{L}_n(\mathbf{\Gamma}_n) = \frac{1}{(2\pi)^{n/2}} |\mathbf{\Gamma}_n|^{-1/2} \exp\left\{ -\frac{1}{2} x_{1:n}^\top \mathbf{\Gamma}_n^{-1} x_{1:n} \right\}.$$

If n is large, computation of

$$|\mathbf{\Gamma}_n| = \det \mathbf{\Gamma}_n \qquad\qquad \mathbf{\Gamma}_n^{-1}$$

may not be straightforward numerically. However, the innovations algorithm can be used to compute the determinant and inverse.

3.5 Estimation

The innovations algorithm returns the quantities

$$\theta_{i,j},\ i = 1, 2, \ldots, n-1,\ j = 1, \ldots, i.$$

If

$$\mathbf{C}_n = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 1 & 0 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 1 & 0 & \cdots & 0 \\ \theta_{3,3} & \theta_{3,2} & \theta_{3,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \theta_{n-1,n-4} & \cdots & 1 \end{bmatrix}$$

Then, as before $X_{1:n} = \mathbf{C}_n(X_{1:n} - \widehat{X}_{1:n})$, with the elements of the $n \times 1$ vector $(X_{1:n} - \widehat{X}_{1:n})$ uncorrelated, with covariance matrix $\mathbf{D}_n = \text{diag}(v_0, v_1, \ldots, v_{n-1})$.

Therefore

$$\mathbf{\Gamma}_n = \mathbf{C}_n \mathbf{D}_n \mathbf{C}_n^\top$$

so that

$$X_{1:n}^\top \mathbf{\Gamma}_n^{-1} X_{1:n} = (X_{1:n} - \widehat{X}_{1:n})^\top \mathbf{D}_n^{-1} (X_{1:n} - \widehat{X}_{1:n})$$

$$= \sum_{j=1}^{n} (X_j - \widehat{X}_j)^2 / v_{j-1}$$

and

$$|\mathbf{\Gamma}_n| = \prod_{j=0}^{n-1} v_j .$$

and thus the innovations algorithm facilitates the Gaussian likelihood computation for any n.

3.5 Estimation

For ARMA(p, q), we saw that

$$\widehat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \widehat{X}_{n+1-j}) & 1 \leqslant n \leqslant m \\ \sum_{j=1}^{p} \phi_j X_{n+1-j} + \sum_{j=1}^{q} \theta_{n,j}(X_{n+1-j} - \widehat{X}_{n+1-j}) & n > m \end{cases}$$

where $m = \max(p, q)$, and

$$v_n = \mathbb{E}[(X_{n+1} - \widehat{X}_{n+1})^2] = \sigma_Z^2 r_n$$

say.

3.5 Estimation

ML estimation requires that the likelihood is maximized, so the innovations algorithm is run for each function evaluation in the iterative optimization algorithm (Newton-Raphson, Quasi-Newton etc.).

The properties of the ML estimators are well-established.

3.5 Estimation

We have that

$$\mathcal{L}(\theta, \phi, \sigma_Z^2) = \left(\frac{1}{2\pi}\right)^{n/2} \left(\frac{1}{\sigma_Z^2}\right)^{n/2} \{R(\phi, \theta)\}^{-1/2} \exp\left\{-\frac{1}{2\sigma_Z^2} S(\phi, \theta)\right\}$$

where

$$R(\phi, \theta) = \prod_{j=1}^{n} r_{j-1} \qquad S(\phi, \theta) = \sum_{j=1}^{n} \frac{(X_j - \widehat{X}_j(\phi, \theta))^2}{r_{j-1}}$$

Note that estimation of σ_Z^2 and (ϕ, θ) can be separated.

3.5 Estimation

Notes:

1. There is no guarantee that the resulting estimates will lead to a causal or invertible process. These constraints can be imposed during the optimization, or a reparameterization to the reciprocal roots can be performed.
2. The identical optimization can be carried out for non-Gaussian processes, leading to maximum pseudo likelihood estimates.

3.5 Estimation

Conditional Likelihood for AR(p): A conditional likelihood approach conditions on the first p data points to facilitate ease of likelihood construction. We have that

$$\begin{aligned} f(y_{1:n}|\phi, \sigma_Z^2) &= f(y_{1:p}|\phi, \sigma_Z^2) f(y_{(p+1):n}|y_{1:p}, \phi, \sigma_Z^2) \\ &= \mathcal{L}_1(\phi, \sigma_Z^2)\mathcal{L}_2(\phi, \sigma_Z^2) \end{aligned}$$

say, where

$$f(y_{(p+1):n}|\phi, \sigma_Z^2) = \prod_{j=p+1}^{n} f(y_j|y_{(j-p-1):(j-1)}, \phi, \sigma_Z^2).$$

The conditional likelihood approach uses only $\mathcal{L}_2(\phi, \sigma_Z^2)$ as the objective function.

3.5 Estimation

Least-squares: A least-squares procedure based on $S(\phi, \theta)$, that is

$$(\widehat{\phi}, \widehat{\theta}) = \arg\min S(\phi, \theta)$$

with

$$\widehat{\sigma}_Z^2 = \frac{S(\widehat{\phi}, \widehat{\theta})}{n - p - q}$$

can be used.

A conditional least-squares approach can also be used for AR(p) models.

3.5 Estimation

In R the main function for ARMA estimation is `arima`:
For example

```
arima(x,order=c(1,0,1))
```

fits an ARMA$(1,1)$ model to the data in `x`.

3.5 Estimation

The library `fArma` contains the `armaFit` function, which provides a number of alternative fitting approaches.

Other functions for AR models include:
`ar`
`ar.burg`
`ar.yw`
`ar.mle`

3.5 Estimation

Example: Lake Huron data (de-trended)

Four potential models:

- AR(1) `arima(Lake.Yt,order=c(1,0,0))`
- AR(2) `arima(Lake.Yt,order=c(2,0,0))`
- MA(1) `arima(Lake.Yt,order=c(0,0,1))`
- ARMA$(1,1)$ `arima(Lake.Yt,order=c(1,0,1))`

Example: Lake Huron data (de-trended)

```
Huron<-scan("https://www.math.mcgill.ca//dstephens//545//Data//huron.txt")
Huron<-Huron+570
n.H<-length(Huron); t.H<-c(1:n.H); x<-c(1875:1972)
Lake.Yt<-residuals(lm(Huron~t.H))
arima(Lake.Yt,order=c(1,0,0))

##
## Call:
## arima(x = Lake.Yt, order = c(1, 0, 0))
##
## Coefficients:
##          ar1  intercept
##       0.7823     0.0794
## s.e.  0.0635     0.3174
##
## sigma^2 estimated as 0.4986:  log likelihood = -105.43,  aic = 216.86
```

Example: Lake Huron data (de-trended)

```
arima(Lake.Yt,order=c(2,0,0))

##
## Call:
## arima(x = Lake.Yt, order = c(2, 0, 0))
##
## Coefficients:
##          ar1      ar2  intercept
##       1.0026  -0.2901     0.0198
## s.e.  0.0977   0.1005     0.2352
##
## sigma^2 estimated as 0.4589:  log likelihood = -101.44,  aic = 210.89
```

3.5 Estimation

Example: Lake Huron data (de-trended)

```
arima(Lake.Yt,order=c(0,0,1))

##
## Call:
## arima(x = Lake.Yt, order = c(0, 0, 1))
##
## Coefficients:
##          ma1  intercept
##       0.7821    -0.0037
## s.e.  0.0646     0.1391
##
## sigma^2 estimated as 0.602:  log likelihood = -114.66,  aic = 235.33
```

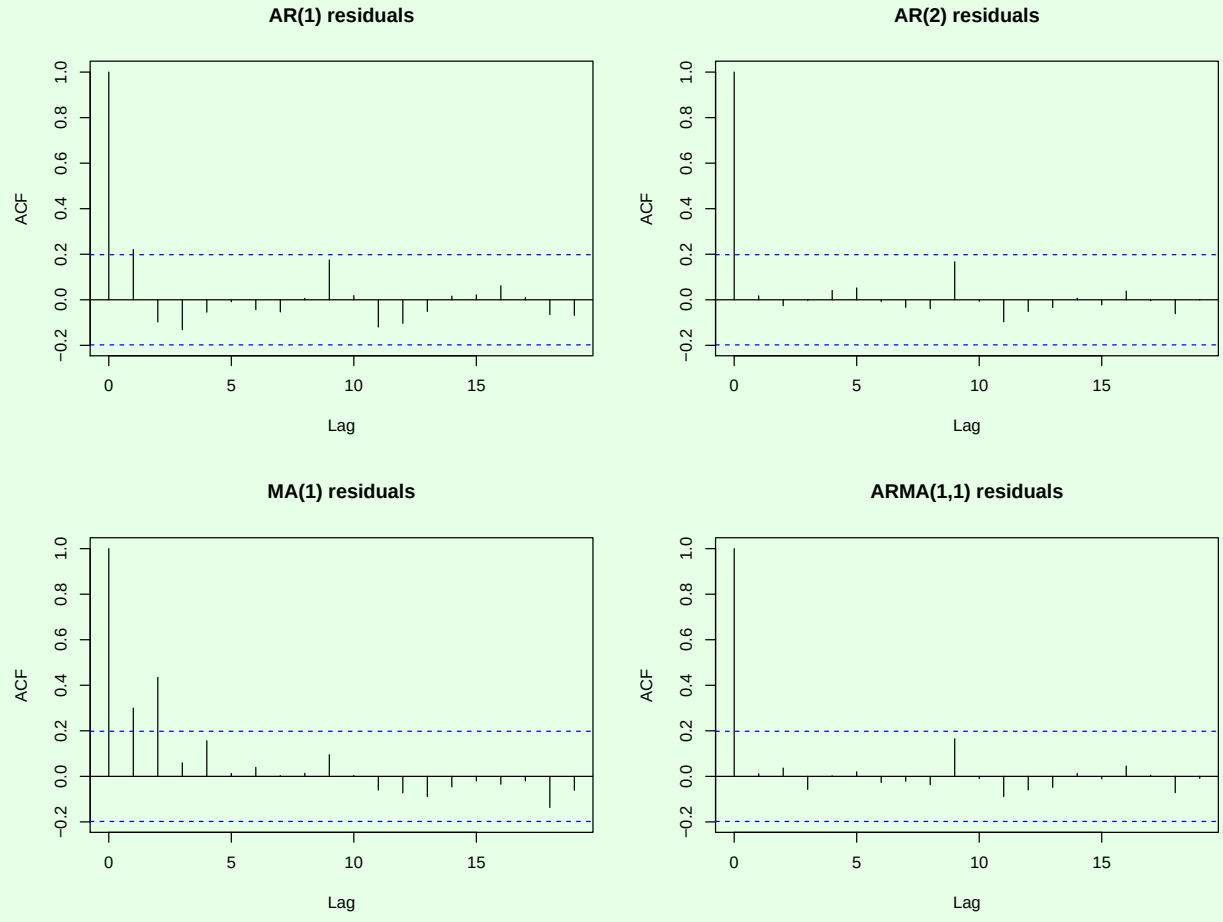
3.5 Estimation

Example: Lake Huron data (de-trended)

```
arima(Lake.Yt,order=c(1,0,1))

##
## Call:
## arima(x = Lake.Yt, order = c(1, 0, 1))
##
## Coefficients:
##          ar1     ma1  intercept
##       0.6508  0.3567     0.0330
## s.e.  0.0948  0.1156     0.2607
##
## sigma^2 estimated as 0.4589:  log likelihood = -101.44,  aic = 210.87
```

Example: Lake Huron data (de-trended)



3.6 Model selection

A series of ARMA models might be fitted, and then compared with reference to different aspects:

- fidelity to the data
- model complexity
- residuals
- forecasting performance

3.6 Model selection

Akaike Information Criterion (AIC)

The AIC model selection method is widely used; it trades off

- fidelity to the observed data (goodness of fit)
- model complexity (number of parameters fitted)

3.6 Model selection

For model M with parameters β yielding likelihood $L_M(\beta)$, the AIC is defined as

$$\mathrm{AIC}(M, \widehat{\beta}) = -2 \log L_M(\widehat{\beta}) + 2d$$

where d is the dimension of β.

For the ARMA(p, q) model, there are $p + q + 1$ parameters $(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q, \sigma_Z^2)$.

3.6 Model selection

For the ARMA(p, q) model,

$$\text{AIC}(M_{p,q}, \widehat{\beta}) = -2\ell_M(\widehat{\phi}_p, \widehat{\theta}_q, S(\widehat{\phi}_p, \widehat{\theta}_q)/n) + 2(p + q + 1)$$

where ℓ_M is the log-likelihood, and

$$\widehat{\sigma}_Z^2 = \frac{1}{n} S(\widehat{\phi}_p, \widehat{\theta}_q)$$

The model with the smallest (ie most negative) AIC is selected.

Notes

- The AIC is a “within sample" criterion, and fits to the observed data. This might not always reflect the objective of the analysis we might be more interested in forecasting performance.
- The AIC is sometimes regarded to give an insufficient penalty for complexity, and chooses models that are too complex.
- As $n \longrightarrow \infty$, the AIC will choose the best approximating model to the data generating model

Corrected AIC

To fix the small sample issue, a correction was suggested:

$$AICC(M_{p,q}, \widehat{\beta}) = -2\ell_M(\widehat{\phi}_p, \widehat{\theta}_q, S(\widehat{\phi}_p, \widehat{\theta}_q)/n) + \frac{2(p+q+1)n}{n-p-q-2}$$

that is the penalty $2d$ is replaced by $2dn/(n-p-q-2)$. This uses a bias-corrected estimator of σ_Z^2.

For large n, the criterion reverts to the AIC.

Bayesian Information Criterion (BIC)

The BIC (or Schwarz criterion) uses a different penalty

$$BIC(M_{p,q}, \widehat{\beta}) = -2\ell_M(\widehat{\phi}_p, \widehat{\theta}_q, S(\widehat{\phi}_p, \widehat{\theta}_q)/n) + (p+q+1)\log n$$

that is the penalty $2d$ is replaced by $d \log n$.

For large n, the criterion correctly selects the true model (provided the true model is one that has been fitted).

3.7 Residual checking

The residual sequence $\{\widehat{W}_t\}$ estimated by

$$\widehat{w}_t = \frac{x_t - \widehat{x}_t}{\sqrt{r_{t-1}}}$$

(that is, the one-step ahead prediction errors standardized by their standard deviation) should have the property that $\{\widehat{W}_t\}$ is an approximate white noise sequence.

3.7 Residual checking

In fact, if the model fitted is the correct model

$$\widehat{W}_t \xrightarrow{\text{m.s.}} Z_t \sim \text{WN}(0, \sigma_Z^2)$$

so the observed residual sequence should be approximately uncorrelated, zero mean and variance σ_Z^2.

We can also compute

$$\hat{R}_t = \frac{\widehat{W}_t}{\hat{\sigma}_Z}$$

which should be a unit variance white noise process.

Chapter 4
Non-Stationary and Seasonal Models

4.1 Non-Stationary and Seasonal Models

Models with non-stationary components can be constructed as straightforward extensions to ARMA models

- Integrated Models: ARIMA
- Seasonal Models: SARIMA

It is also possible to specify time-varying or stochastic mean or variance models.

4.2 ARIMA Models

If d is a non-negative integer, then $\{X_t\}$ is an ARIMA(p, d, q) process if $\{Y_t\}$ defined by

$$Y_t = (1 - B)^d X_t$$

is a (causal) ARMA(p, q) process, that is

$$\Phi(B)(1 - B)^d X_t = \Theta(B) Z_t$$

where $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$; $\{X_t\}$ is stationary only if $d = 0$.

The ARIMA(p, d, q) process can be made stationary by dth order differencing.

Example: ARIMA$(1,1,0)$ vs ARIMA$(2,0,0)$

If $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$:

- **Stationary** $\text{ARIMA}(2,0,0) \equiv \text{ARMA}(2,0) \equiv \text{AR}(2)$

$$(1 - 0.6B)(1 - 0.99B)X_t = Z_t$$

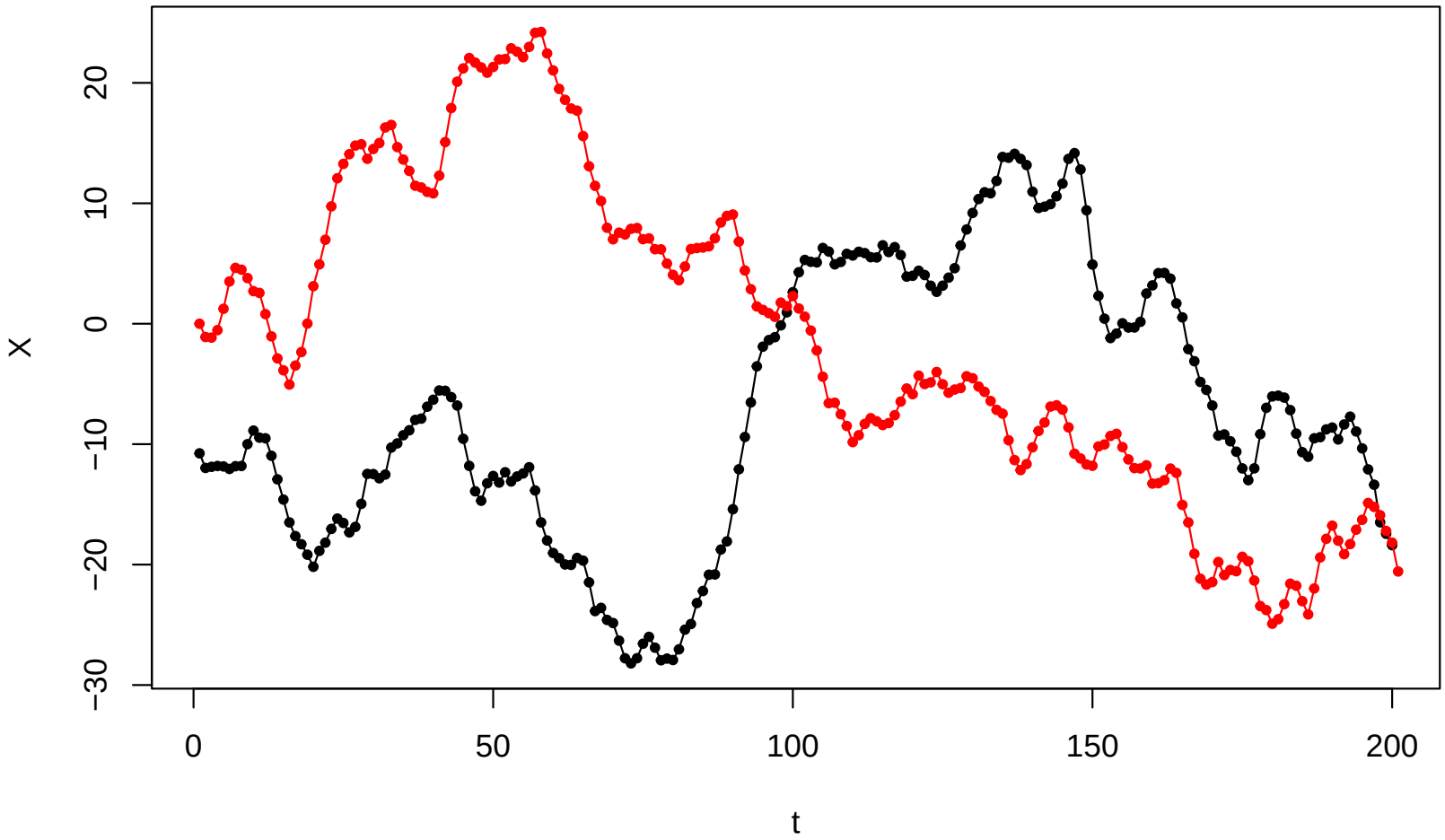
- **Non-stationary** $\text{ARIMA}(1,1,0)$

$$(1 - 0.6B)(1 - B)X_t = Z_t$$

4.2 ARIMA Models

```
n<-200;set.seed(164664)
r1<-0.6;r2<-0.99
x1<-arima.sim(n,model=list(ar=c(r1+r2,-r1*r2),order=c(2,0,0)))
x2<-arima.sim(n,model=list(ar=c(r1),order=c(1,1,0)))
par(mar=c(4,4,3,2))
plot(x1,xlab=expression(t),type='p',pch=19,cex=0.6,ylim=range(c(x1,x2)),ylab='X')
lines(x1);
points(x2,pch=19,cex=0.6,col='red');lines(x2,col="red")
```

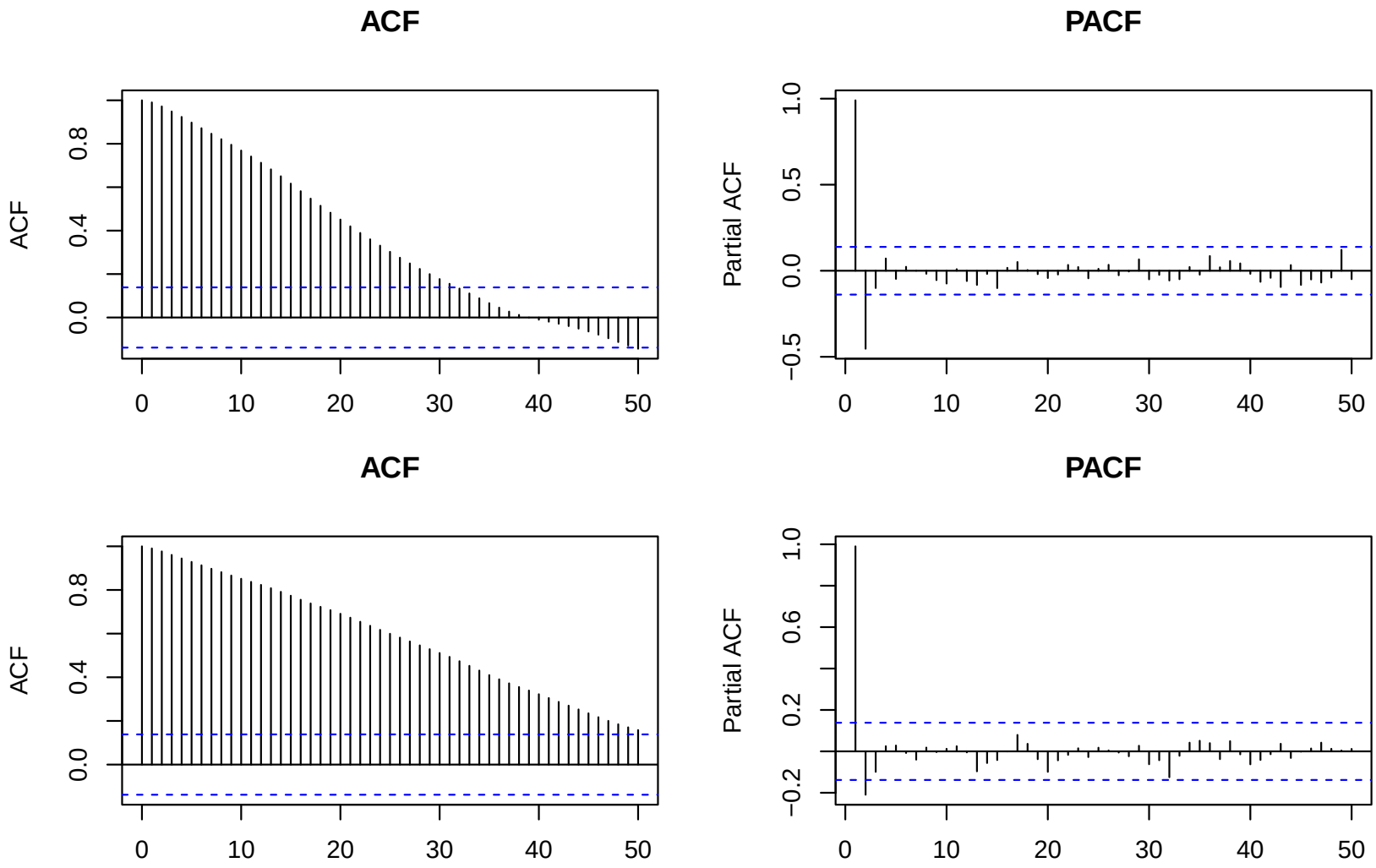
4.2 ARIMA Models



4.2 ARIMA Models

```
par(mfrow=c(2,2),mar=c(3,4,3,2))
acf(x1,main="ACF",lag.max=50)
pacf(x1,main="PACF",lag.max=50)
acf(x2,main="ACF",lag.max=50)
pacf(x2,main="PACF",lag.max=50)
```

4.2 ARIMA Models



Consider the stationary AR(2) model with complex reciprocal roots

$$(1 - \xi B)(1 - \overline{\xi} B) X_t = Z_t$$

and the corresponding AR polynomial

$$\Phi(z) = (1 - \xi z)(1 - \overline{\xi} z) = 1 - 2\mathrm{Re}(\xi) z + |\xi|^2 z^2$$

Suppose that $\xi = \exp\{-i\omega\}/r$, where $r > 1$ and $-\pi < \omega < \pi$. Then

$$\Phi(z) = 1 - \left(\frac{2}{r}\cos\omega\right) z + \left(\frac{1}{r^2}\right) z^2.$$

It follows that

$$\rho_X(h) = \left(\frac{1}{r}\right)^h \frac{\sin(h\omega + \lambda)}{\sin\lambda} \qquad h = 0, \pm 1, \pm 2, \ldots$$

where

$$\lambda = \arctan\left[\left(\frac{r^2+1}{r^2-1}\right)\tan\omega\right].$$

In the limit as $r \downarrow 1$

$$\rho_X(h) \longrightarrow \cos(h\omega).$$

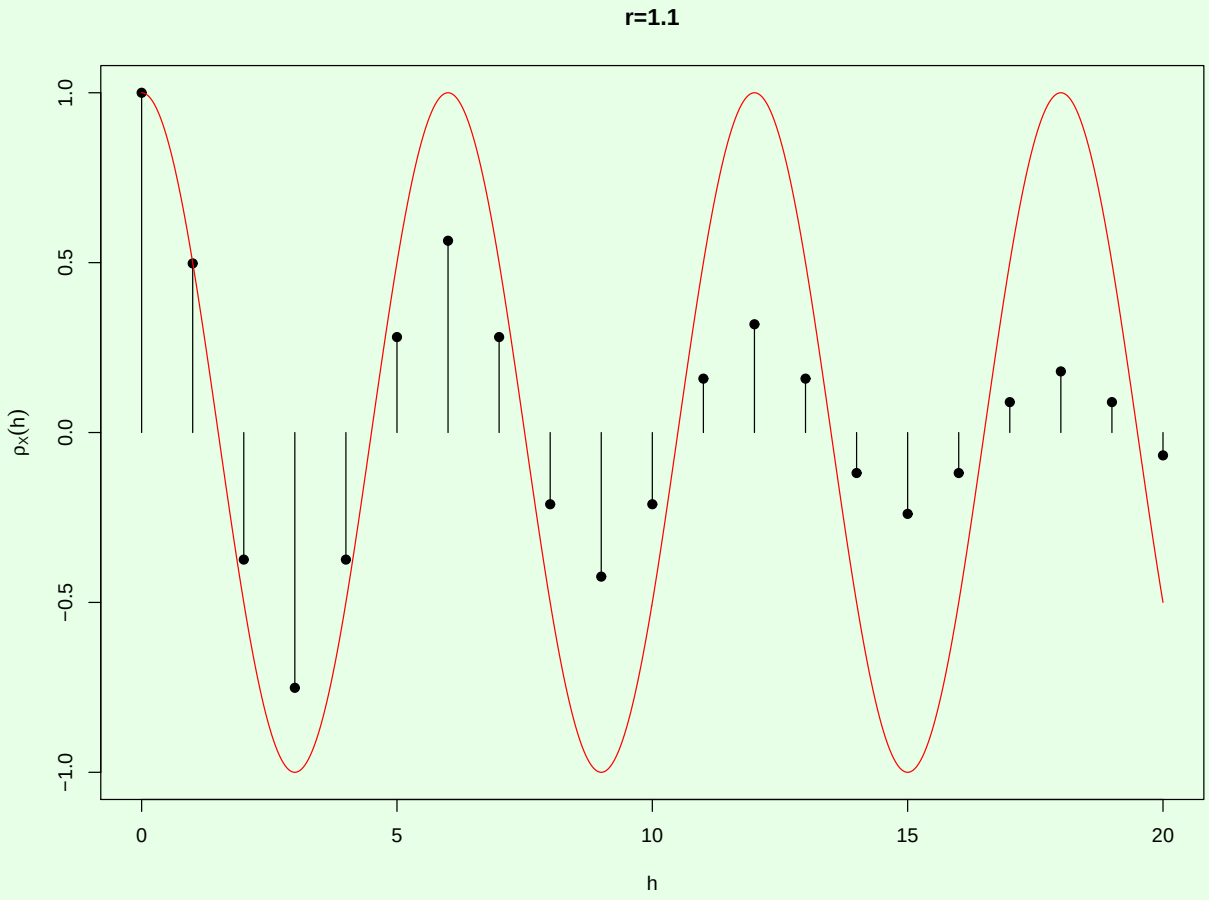
4.2 ARIMA Models

We examine this model as

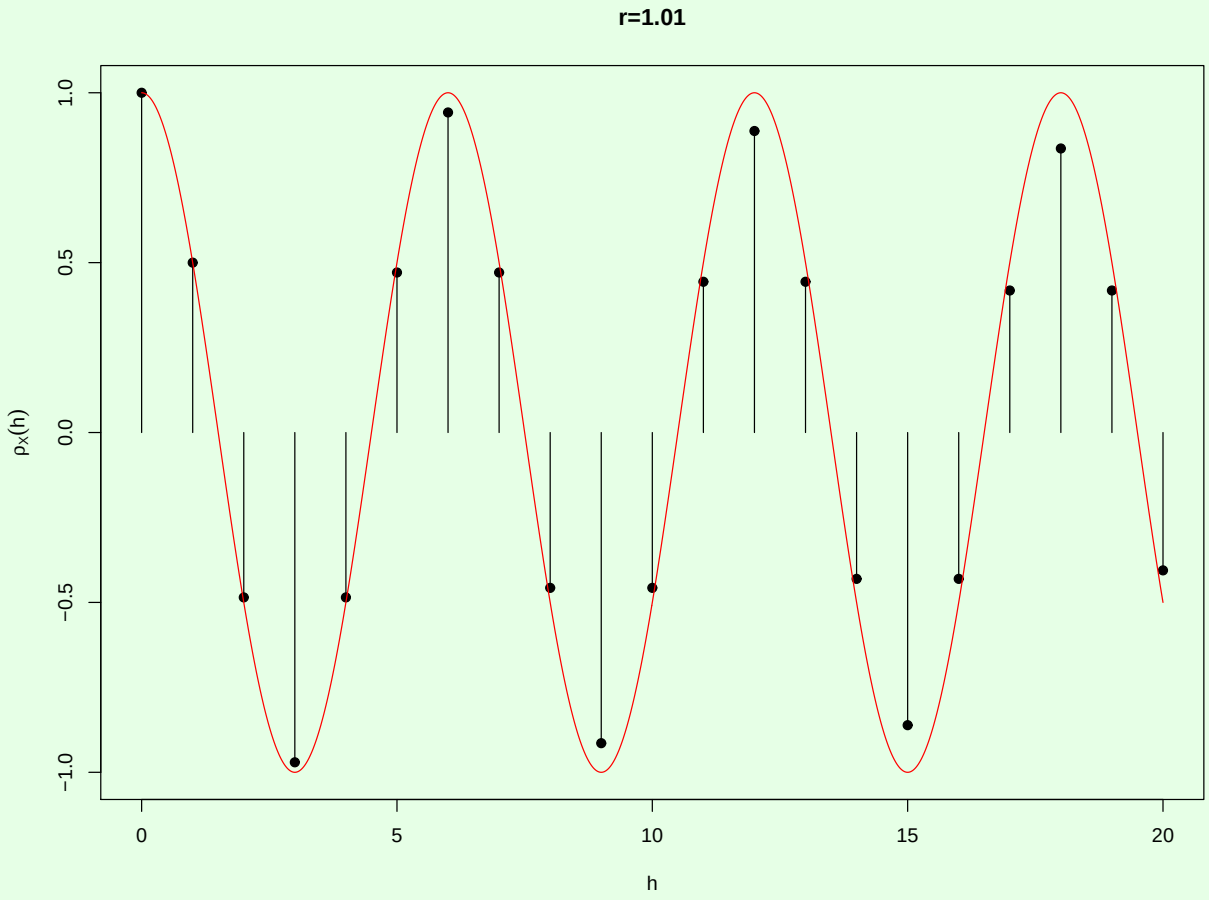
- r takes values tending to 1 from above
- ω fixed.

and compare with the pure $\cos$ function.

Example: $r = 1.1, \omega = \pi/3$

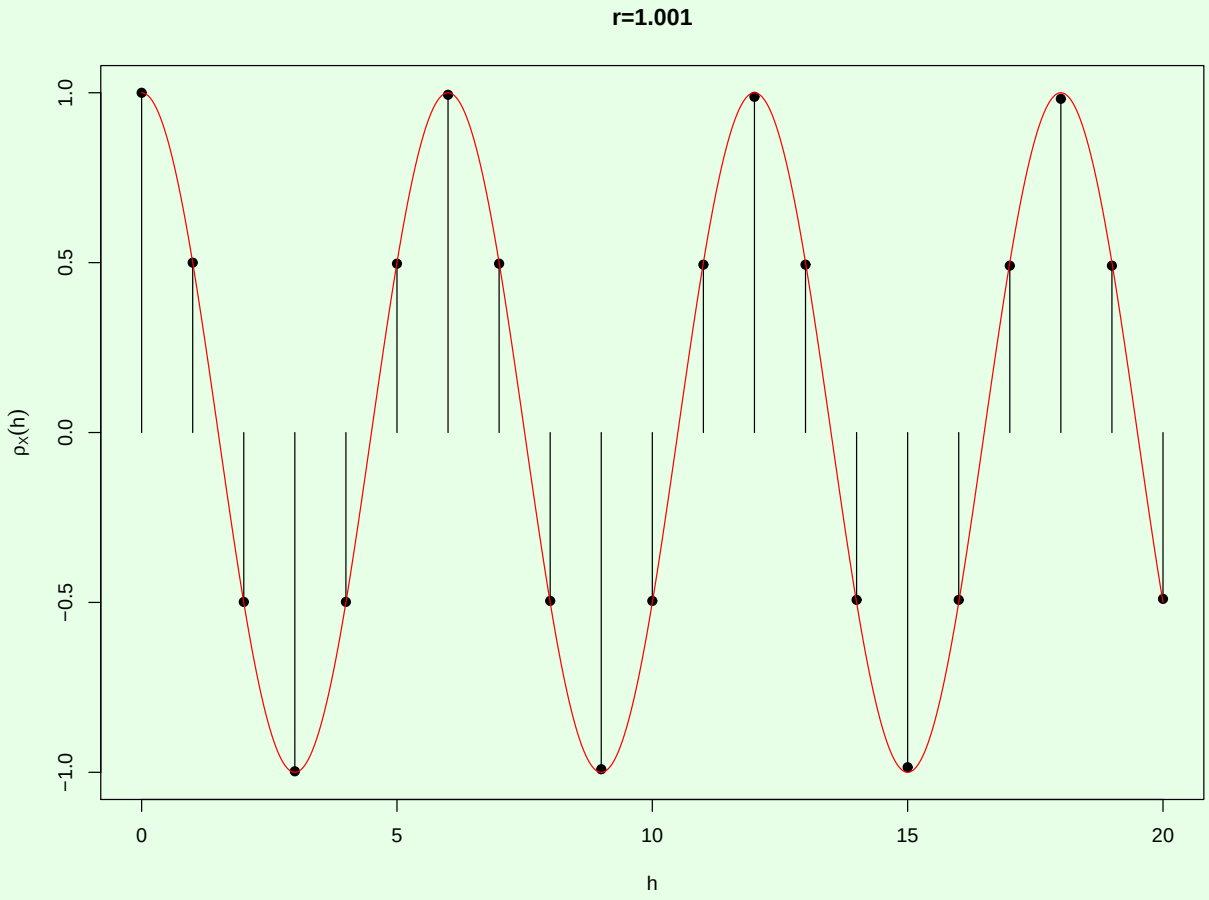


Example: $r = 1.01, \omega = \pi/3$

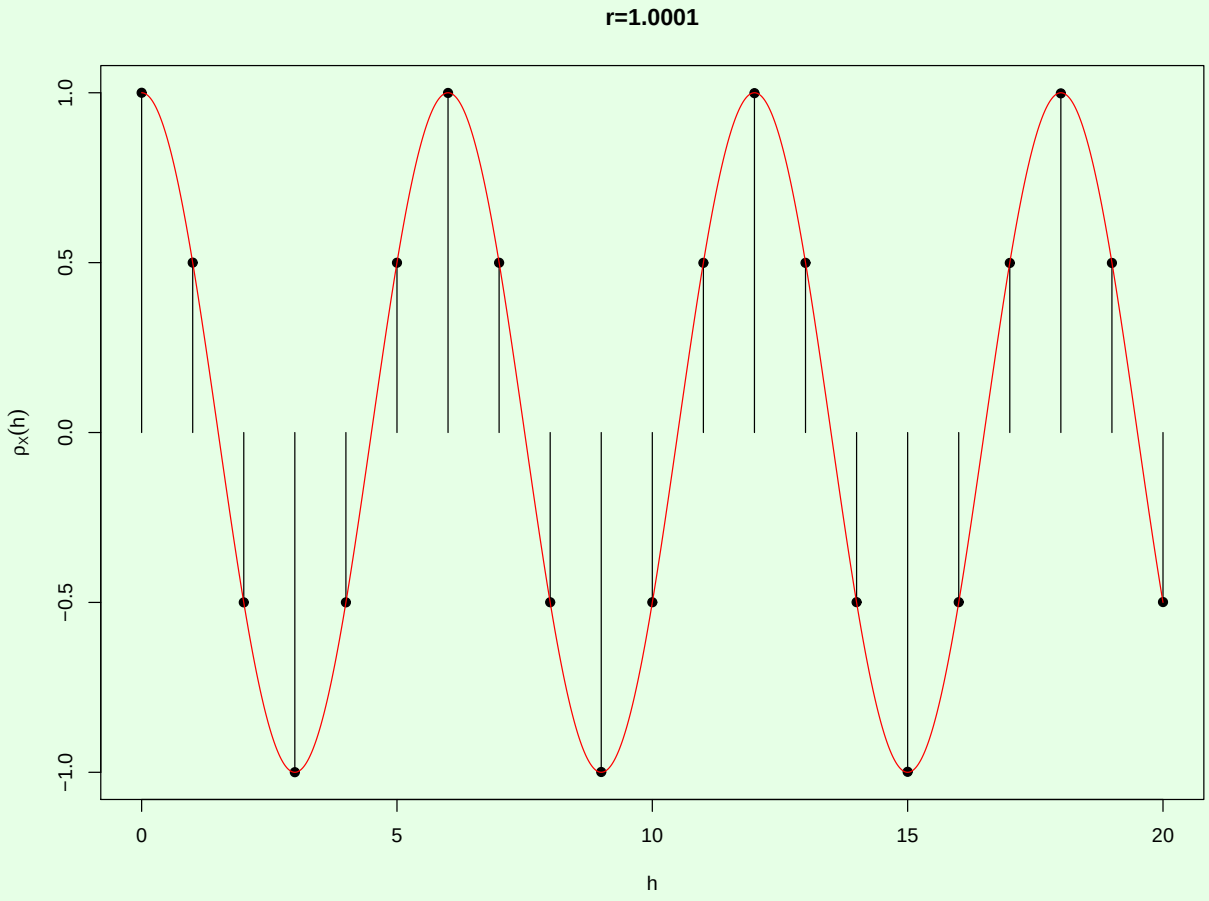


4.2 ARIMA Models

Example: $r = 1.001, \omega = \pi/3$

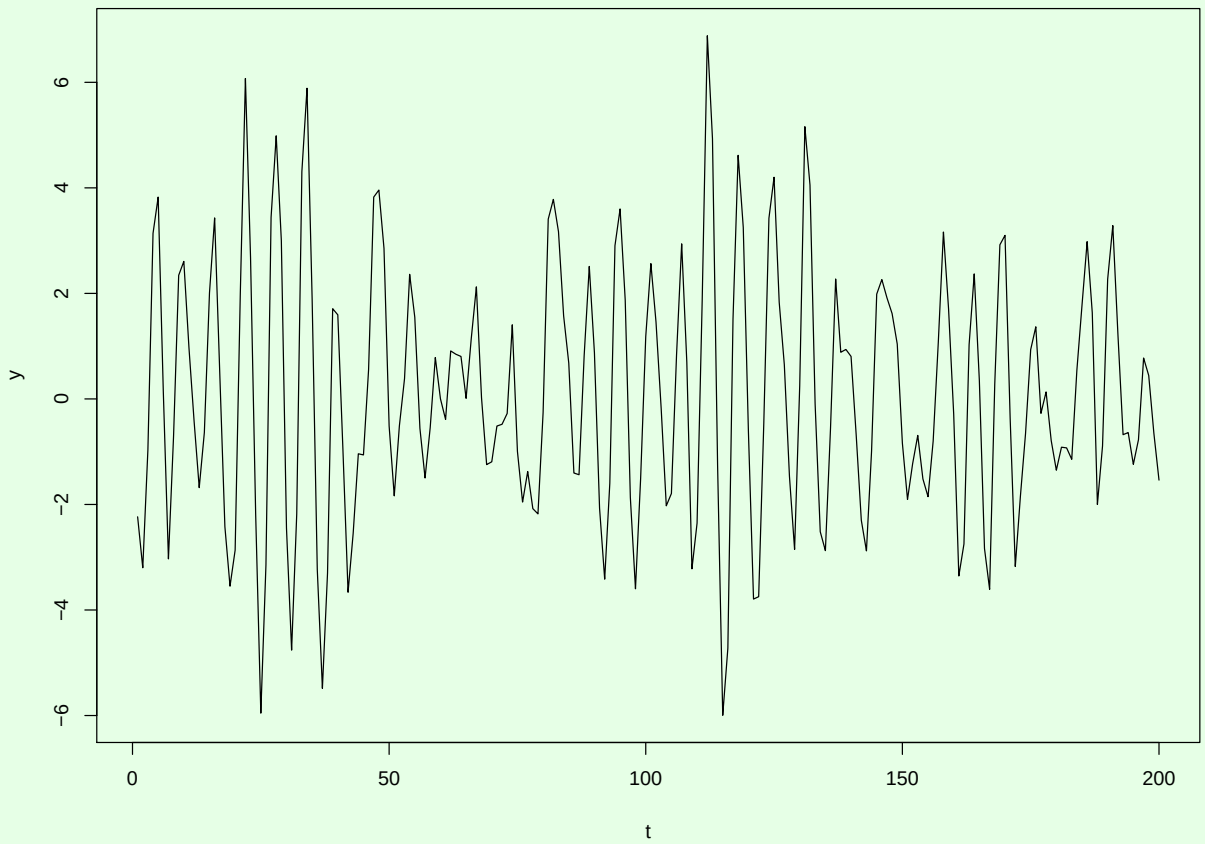


Example: $r = 1.0001, \omega = \pi/3$

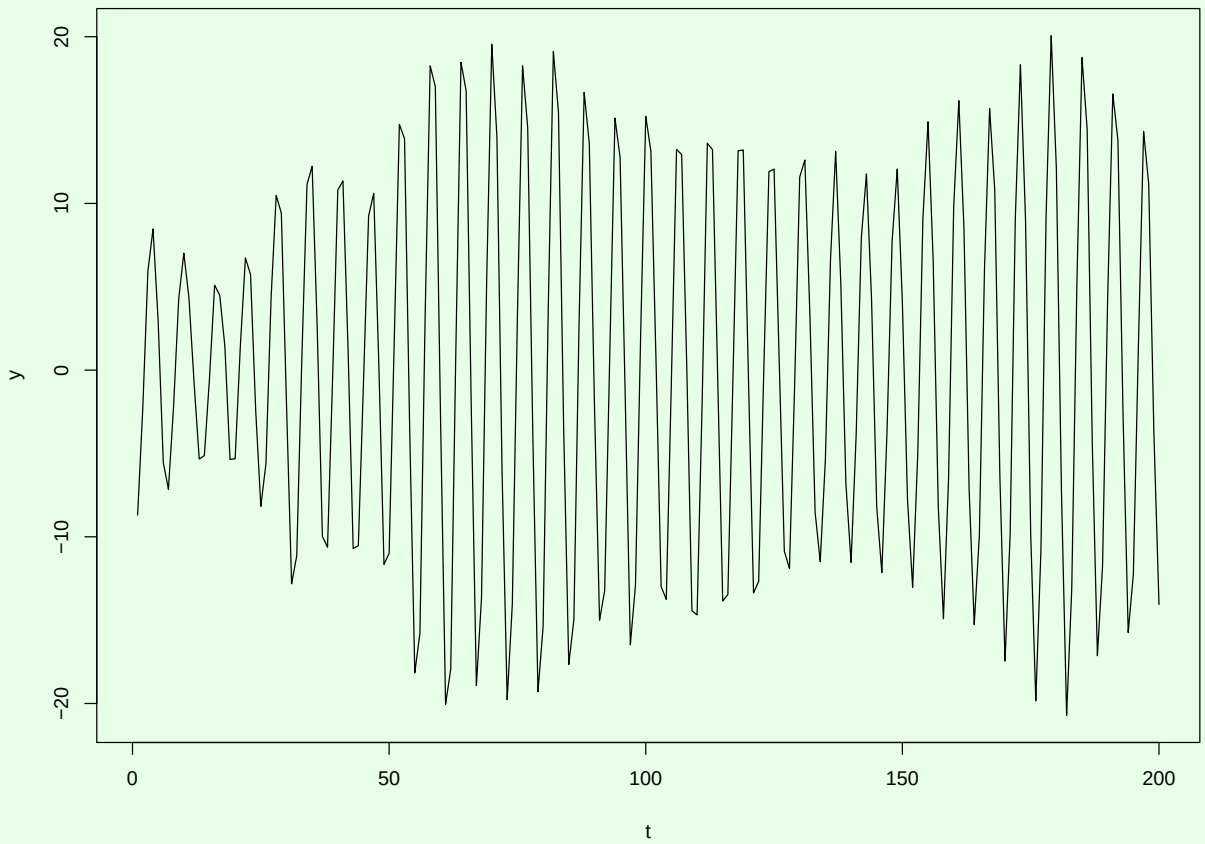


4.2 ARIMA Models

Example: $r = 1.1, \omega = \pi/3$



Example: $r = 1.001, \omega = \pi/3$



This demonstrates that we can construct non-stationary processes with different behaviour from the random walk model

$$(1 - B)X_t = Z_t$$

that is, we can construct a periodic non-stationary model by choosing

$$|\xi| \longrightarrow 1$$

with ξ having an imaginary part.

4.3 Unit Roots

The presence of a unit root of the AR polynomial equation

$$\Phi(z) = 0$$

that is, a solution where $|\xi| = 1$, renders the resulting AR or ARMA process non-stationary.

Therefore it is crucial to be able to identify processes where a unit root is present, and to distinguish this from a root near to, but not equal to, one in modulus.

4.3 Unit Roots

The Dickey-Fuller Test: Suppose $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and

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

with $|\phi| < 1$. Then, by standard asymptotic theory, ML or moment-based estimators $\widehat{\phi}$ satisfy

$$\sqrt{n}(\widehat{\phi} - \phi) \xrightarrow{d} \mathcal{N}(0, 1 - \phi^2)$$

as $n \longrightarrow \infty$.

If $|\phi| = 1$, the asymptotic result does not hold.

4.3 Unit Roots

We may study the distribution of the statistic

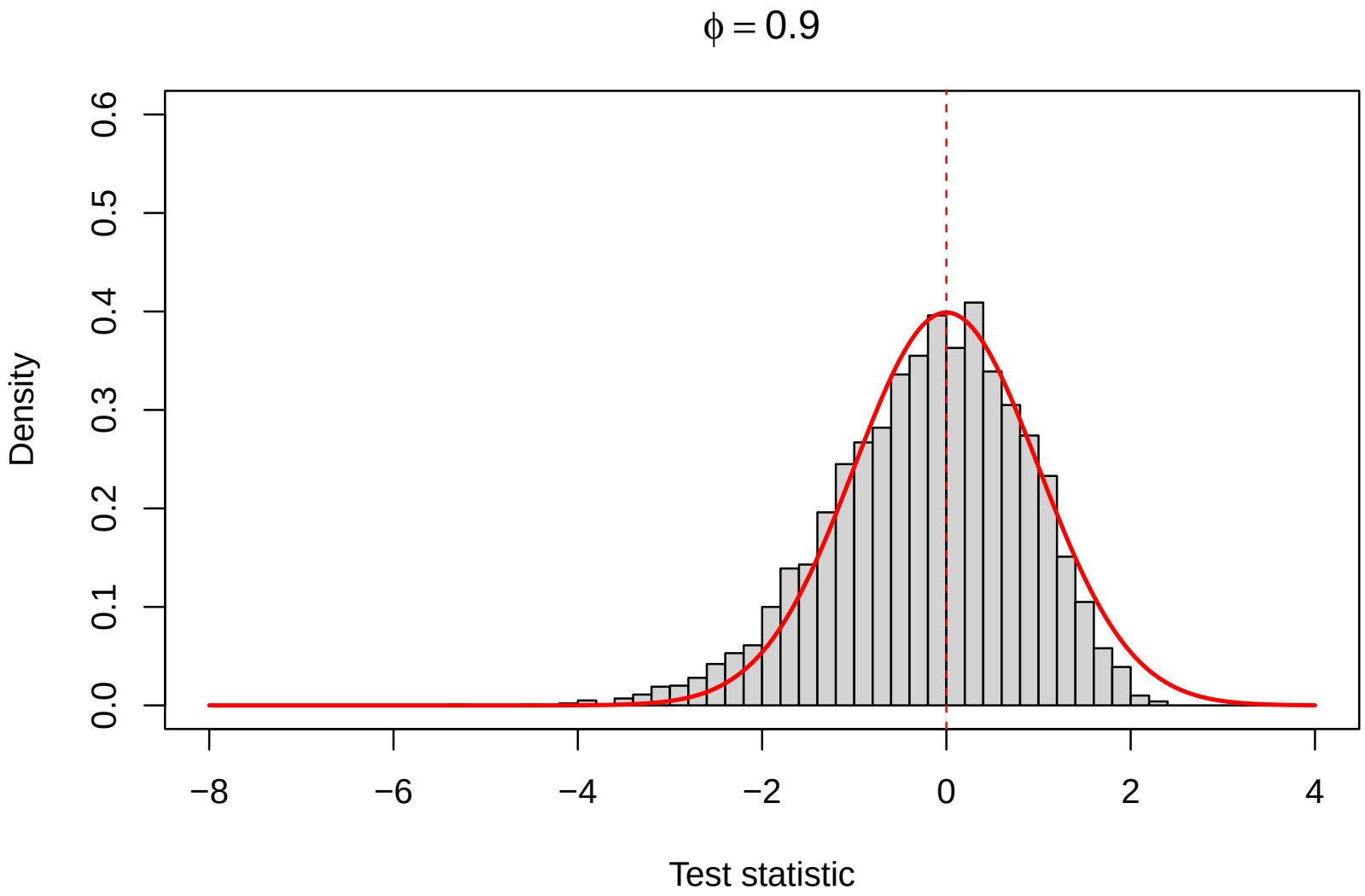
$$\frac{\sqrt{n}(\widehat{\phi} - \phi)}{\sqrt{1 - \phi^2}}$$

in simulation for any value of ϕ, which should be standard Normally distributed when n is large.

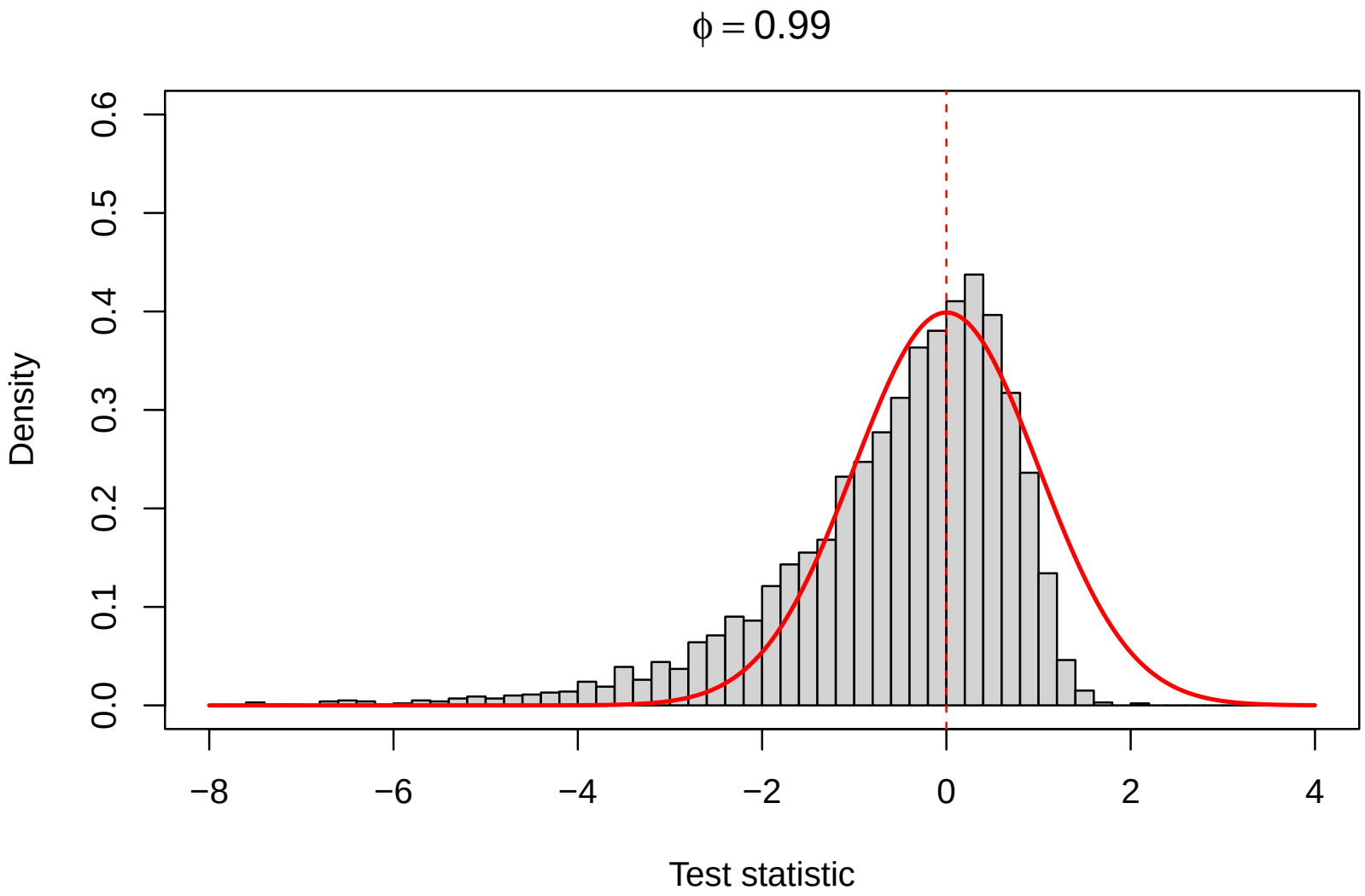
In the following simulations, we use $n = 500$ for

- $\phi = 0.9$
- $\phi = 0.99$
- $\phi = 0.999$

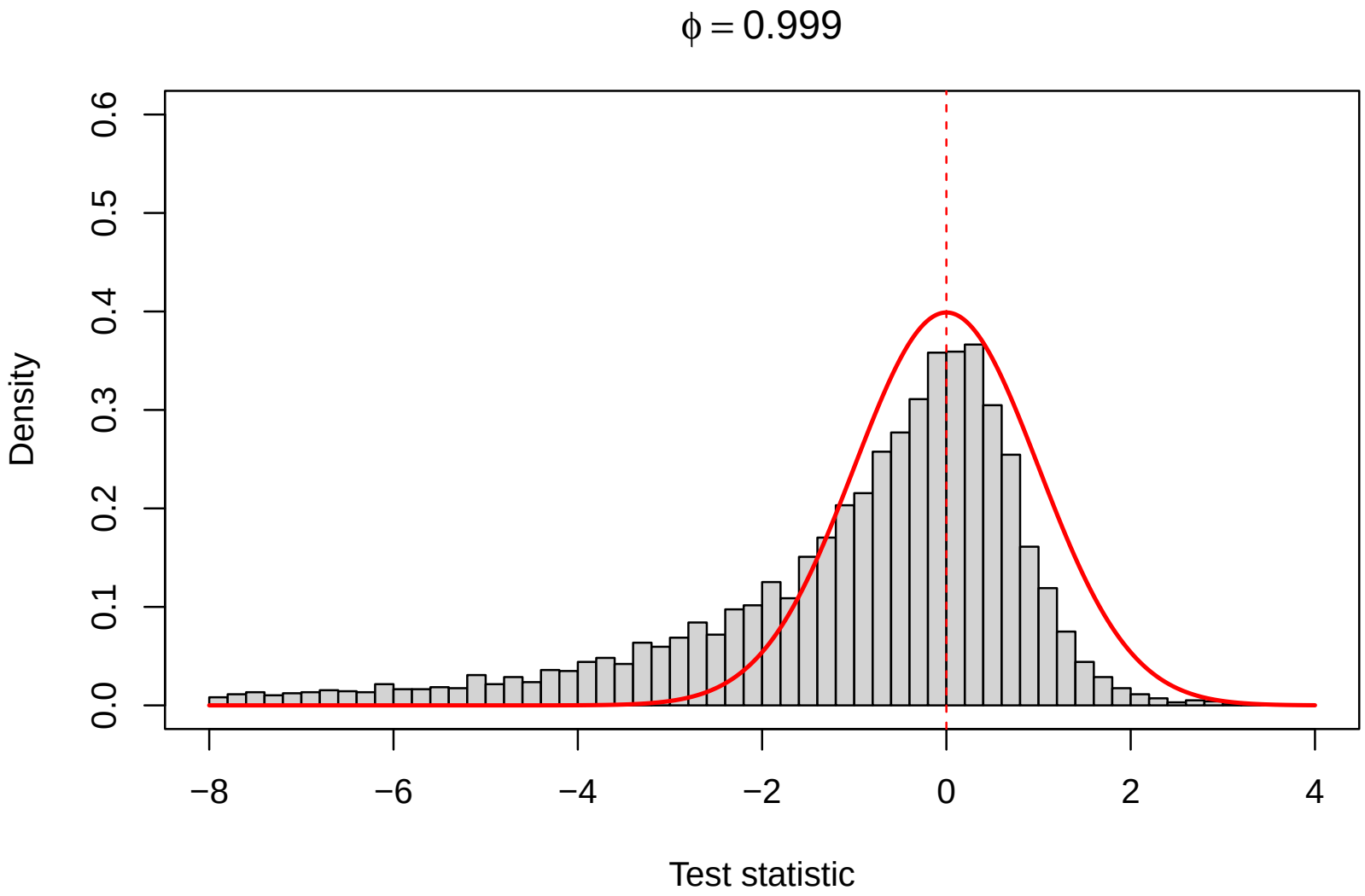
4.3 Unit Roots



4.3 Unit Roots



4.3 Unit Roots



4.3 Unit Roots

The Dickey-Fuller Test: Suppose $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and

$$X_t = \mu + \phi(X_{t-1} - \mu) + Z_t.$$

Then

$$\begin{aligned}(1-B)X_t = X_t - X_{t-1} &= \mu(1-\phi) + (\phi - 1)X_{t-1} + Z_t \\ &= \phi_0^* + \phi_1^* X_{t-1} + Z_t\end{aligned}$$

where

$$\phi_0^* = \mu(1-\phi) \qquad\qquad \phi_1^* = \phi - 1.$$

can be estimated by OLS, regressing X_t on X_{t-1}.

4.3 Unit Roots

To test the hypothesis

$$H_0 : \phi_1^* = 0$$

which is equivalent to

$$H_0 : \phi_1 = 1$$

the Dickey-Fuller test statistic

$$t = \frac{\widehat{\phi}_1^*}{\text{e.s.e}(\widehat{\phi}_1^*)}$$

can be used. Numerator and denominator are available from the linear regression.

4.3 Unit Roots

However, if H_0 is true, and $\{X_t\}$ is non-stationary, the null distribution is non-standard.

In this simple case, the form of the null distribution can be computed, but it is not available in closed form. It is straightforward to compute using simulation.

4.3 Unit Roots

```
n<-10
Zt<-rnorm(n)
Xt<-cumsum(Zt)
X<-Xt[1:(n-1)]
tvals<-summary(lm(diff(Xt)~-1+X))$coef
tvals

##    Estimate Std. Error  t value  Pr(>|t|)
## X 0.1527851  0.1101742 1.386759 0.2029315

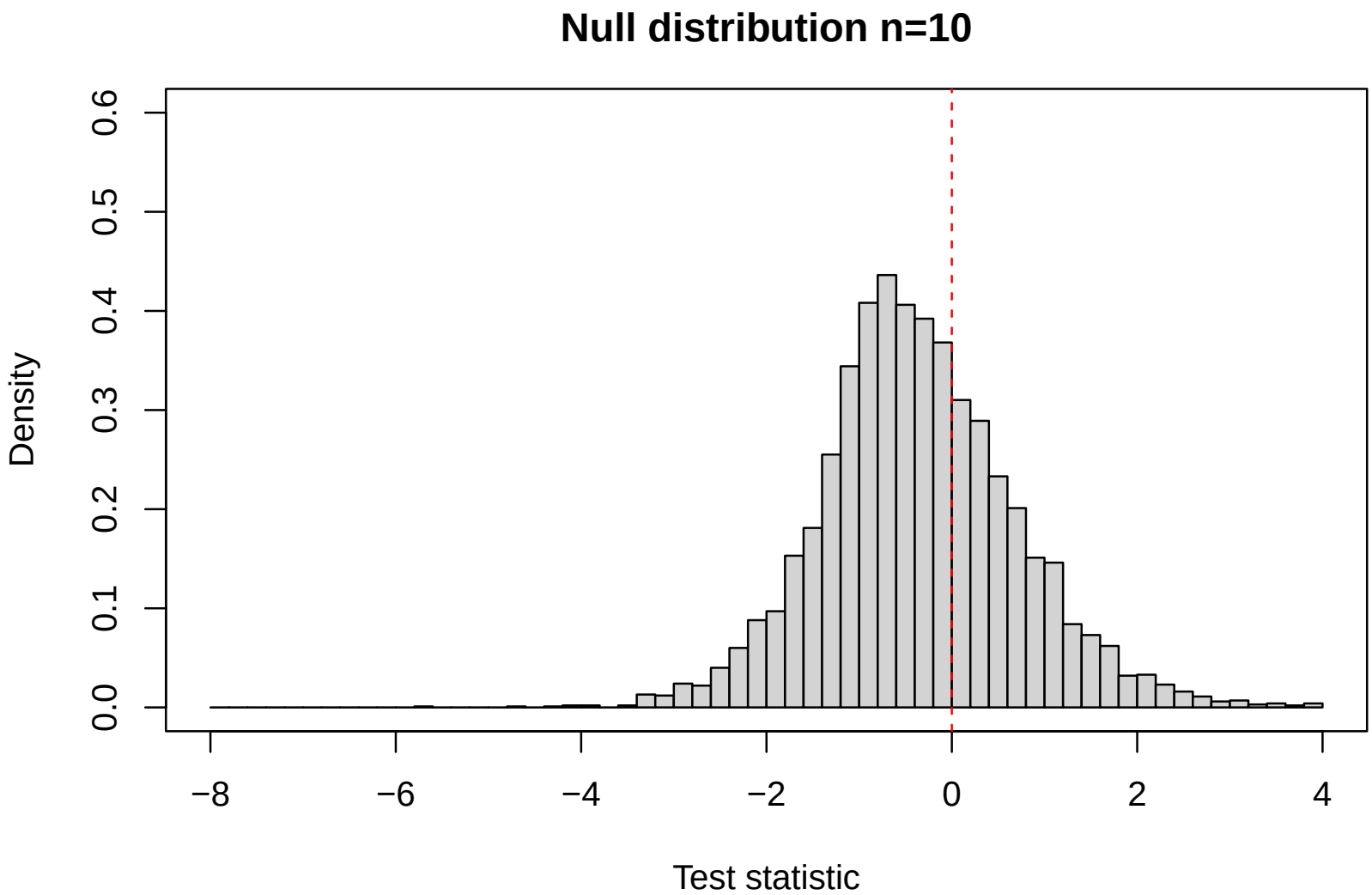
tstat<-(tvals[1,1]-1)/tvals[1,2]
tstat

## [1] -7.689774
```

4.3 Unit Roots

```
nreps<-5000  #5000 replications
n<-10
tstat10<-rep(0,nreps)
for(irep in 1:nreps){
    Zt<-rnorm(n)
    Xt<-cumsum(Zt)
    X<-Xt[1:(n-1)]
    tvals<-summary(lm(diff(Xt)~-1+X))$coef
    tstat<-tvals[1,1]/tvals[1,2]
    tstat10[irep]<-tstat
}
tstat10<-tstat10[tstat10 > -8 & tstat10 < 4]
par(mar=c(4,4,3,2))
hist(tstat10,freq=FALSE,breaks=seq(-8,4,by=0.2),ylim=range(0,0.6),
        main='Null distribution n=10',xlab='Test statistic')
box()
abline(v=0,lty=2,col='red')
```

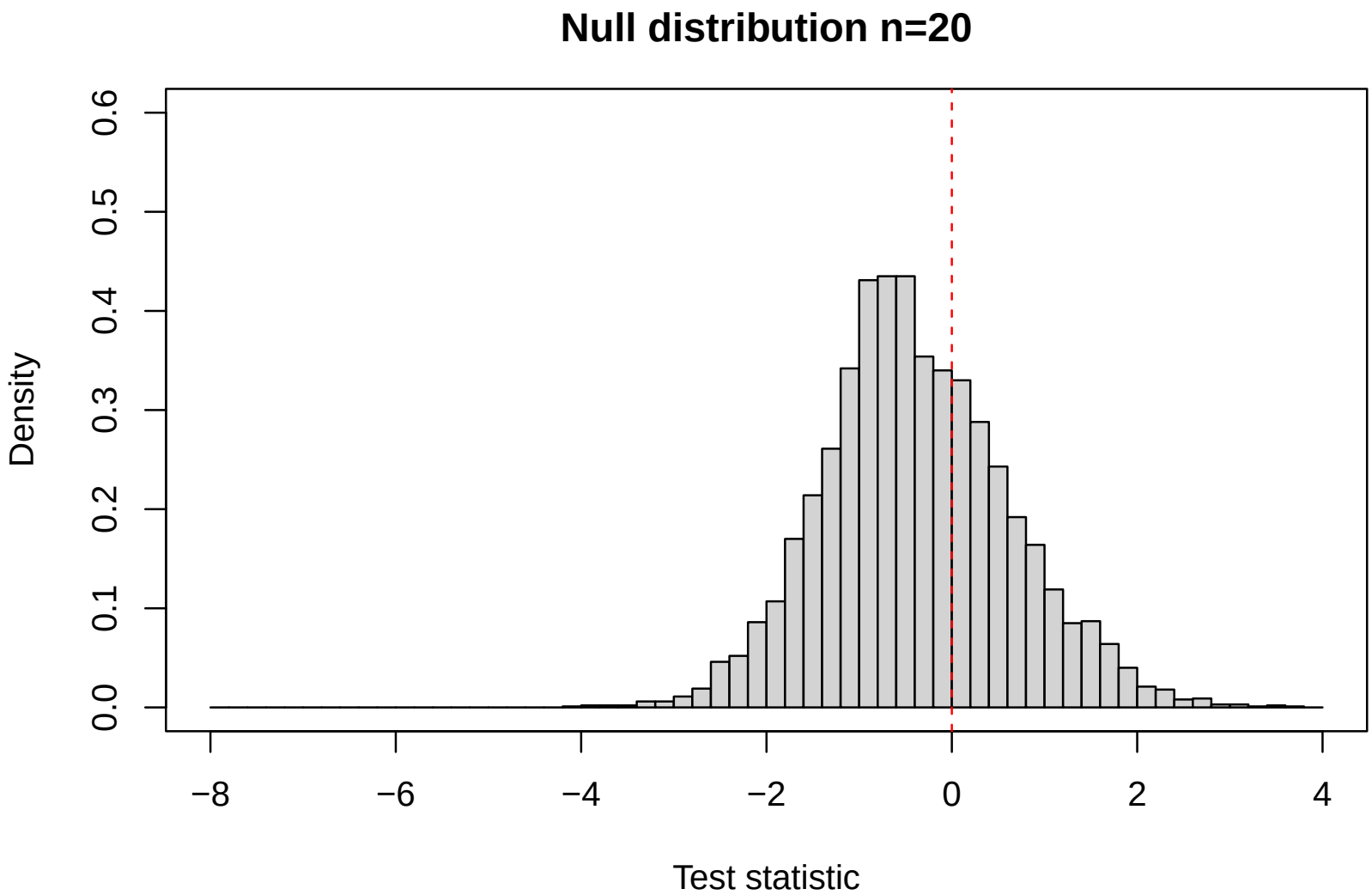
4.3 Unit Roots



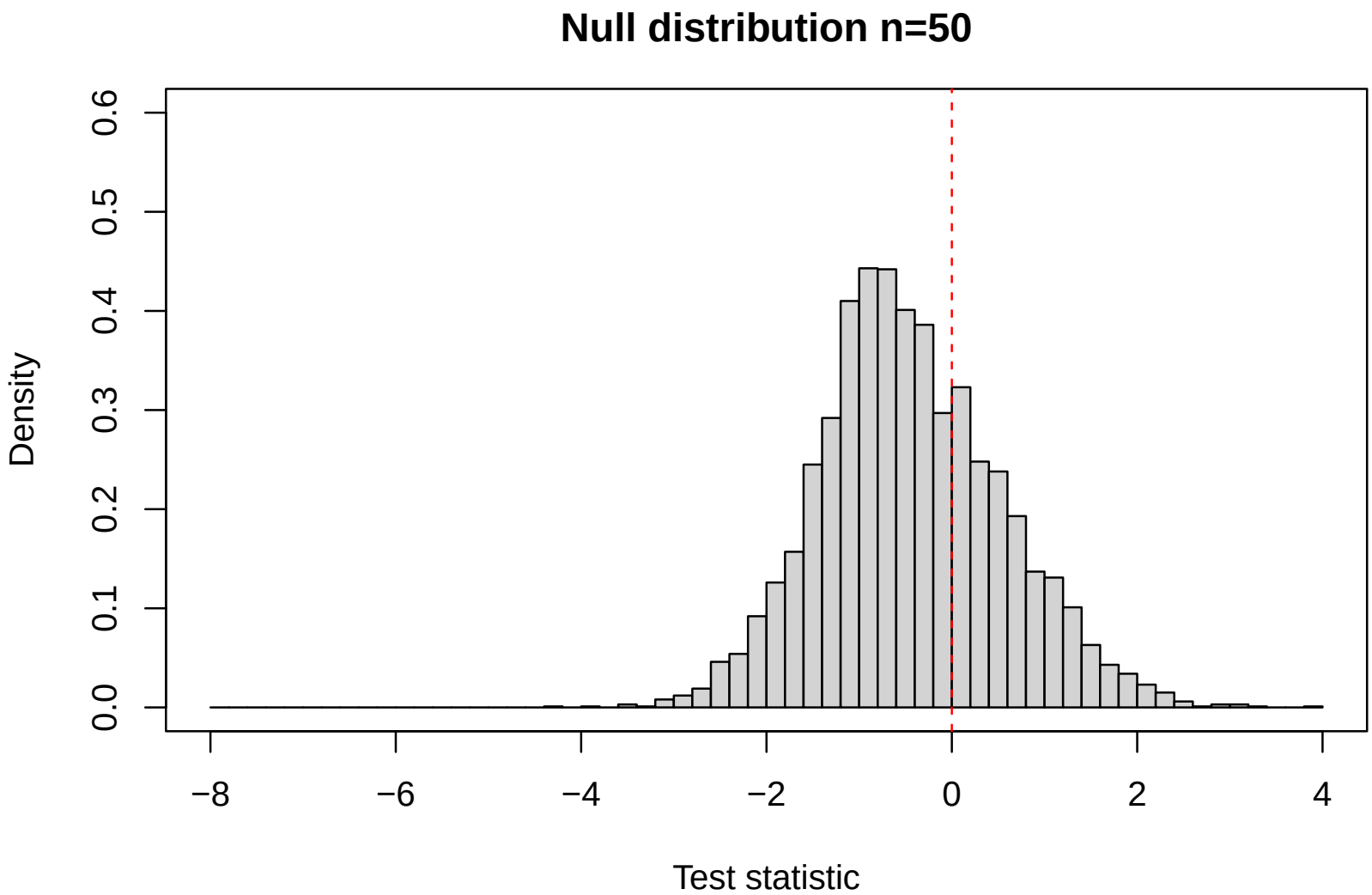
4.3 Unit Roots

```
nreps<-5000  #5000 replications
n<-20
tstat20<-rep(0,nreps)
for(irep in 1:nreps){
    Zt<-rnorm(n)
    Xt<-cumsum(Zt)
    X<-Xt[1:(n-1)]
    tvals<-summary(lm(diff(Xt)~-1+X))$coef
    tstat<-tvals[1,1]/tvals[1,2]
    tstat20[irep]<-tstat
}
tstat20<-tstat20[tstat20 > -8 & tstat20 < 4]
par(mar=c(4,4,3,2))
hist(tstat20,freq=FALSE,breaks=seq(-8,4,by=0.2),ylim=range(0,0.6),
        main='Null distribution n=20',xlab='Test statistic')
box()
abline(v=0,lty=2,col='red')
```

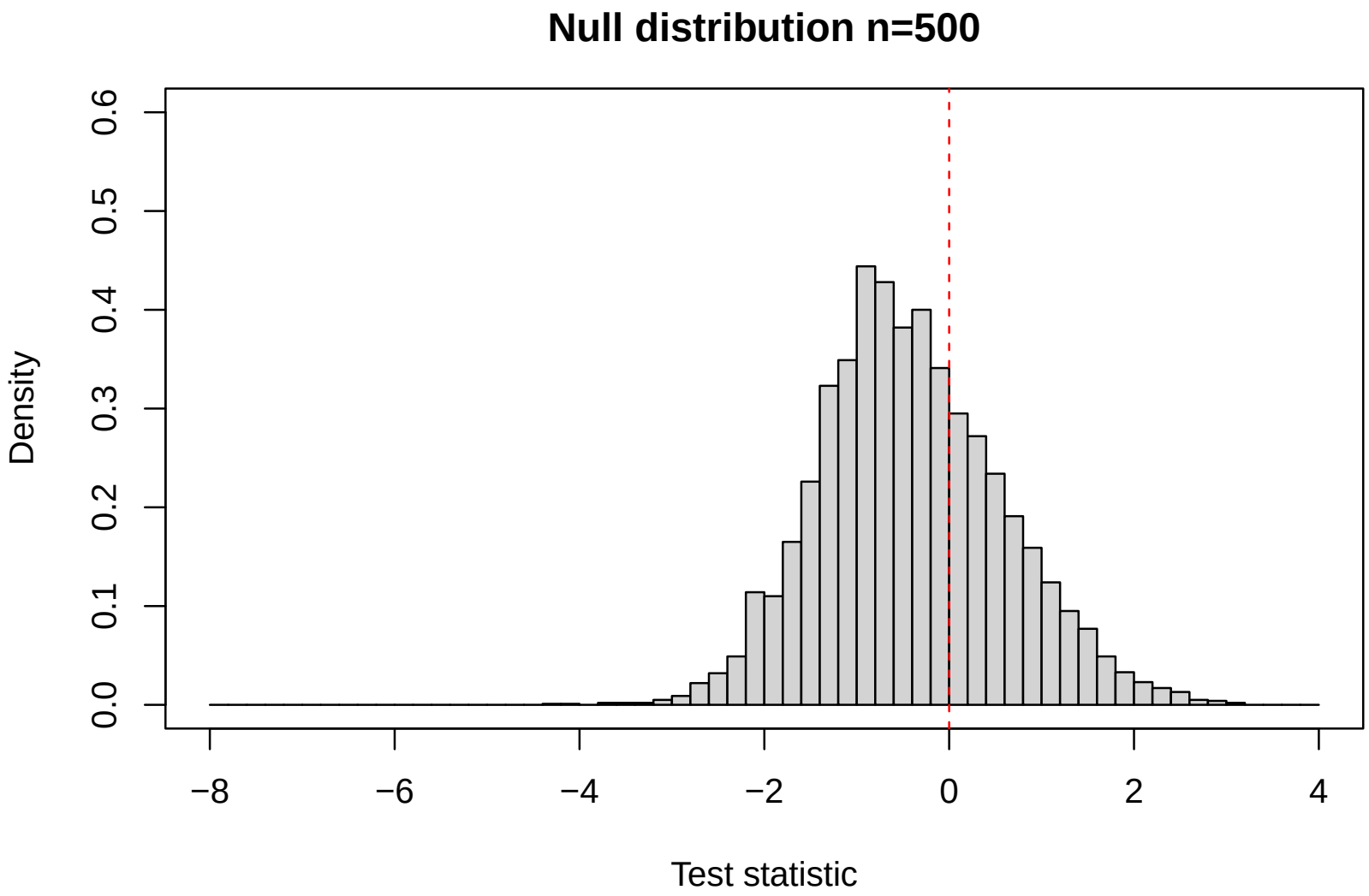
4.3 Unit Roots



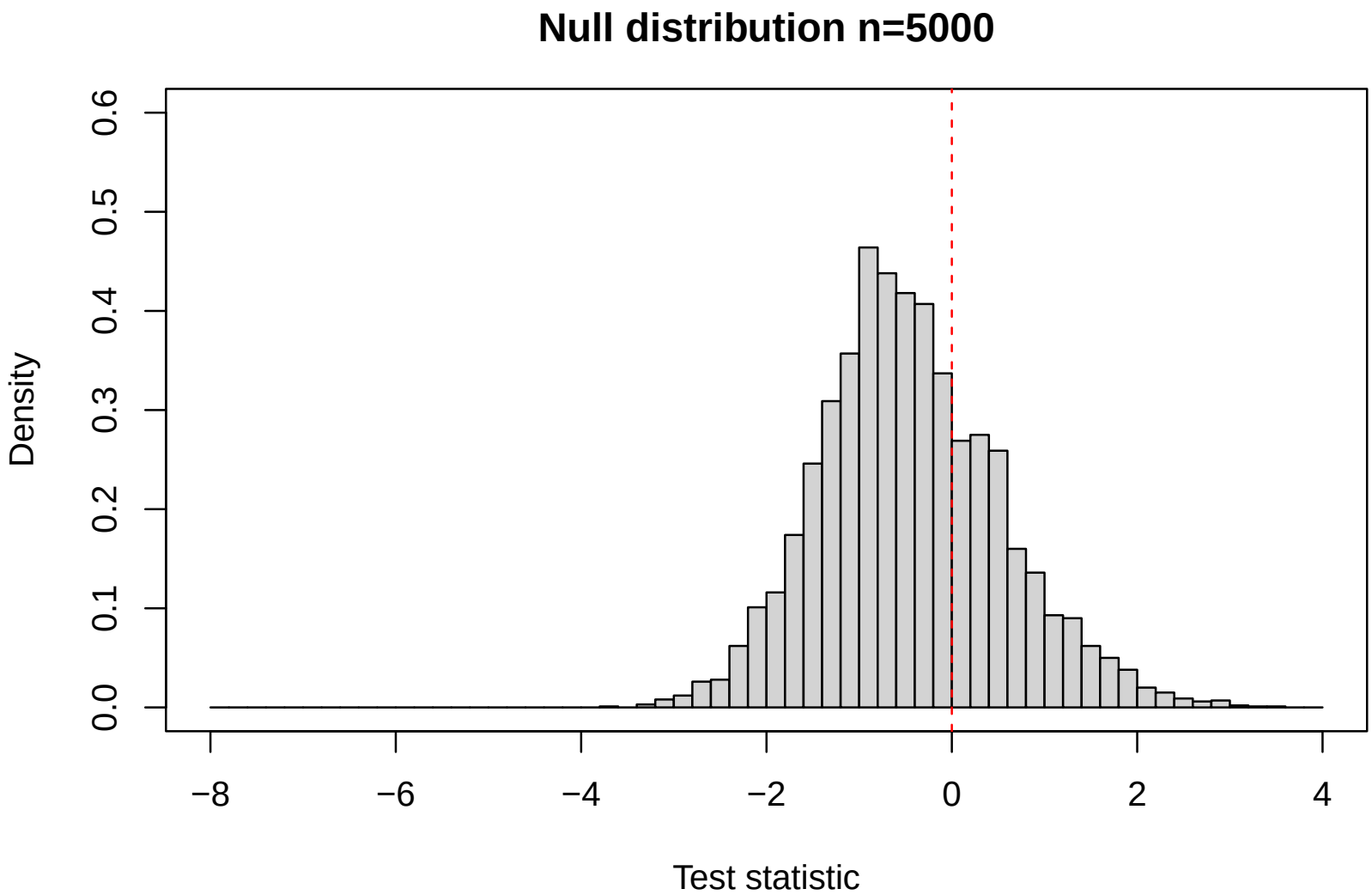
4.3 Unit Roots



4.3 Unit Roots



4.3 Unit Roots



4.3 Unit Roots

The test can be extended to AR(p) processes: if

$$X_t - \mu = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu) + Z_t$$

then for $p \geqslant 2$,

$$(1 - B)X_t = \phi_0^* + \phi_1^* X_{t-1} + \sum_{j=2}^{p} \phi_i^* (X_{t-j+1} - X_{t-j}) + Z_t$$

where

$$\phi_0^* = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$$

and

$$\phi_1^* = \sum_{j=1}^{p} \phi_j - 1 \qquad \phi_j^* = -\sum_{k=j}^{p} \phi_k \qquad j = 2, \ldots, p.$$

4.3 Unit Roots

If $\{X_t\} \sim \text{AR}(p)$ with one unit root, then

$$(1 - B)X_t \sim \text{AR}(p - 1)$$

therefore a unit root test can be carried out by examining the hypothesis

$$H_0 \ : \ \phi_1^* = 0$$

by fitting a regression of $(1 - B)X_t$ on

$$1, X_{t-1}, X_{t-1} - X_{t-2}, \ldots, X_{t-p+1} - X_{t-p}$$

The null distribution of the test statistic changes with p.

4.3 Unit Roots

The general procedure is the Augmented Dickey-Fuller test.

R Code: The library `tseries` contains the function `adf.test`

4.3 Unit Roots

```
library(tseries)

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

x <- rnorm(1000)
adf.test(x,k=0)

## Warning in adf.test(x, k = 0):  p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  x
## Dickey-Fuller = -33.086, Lag order = 0, p-value = 0.01
## alternative hypothesis: stationary

adf.test(cumsum(x),k=0)

##
##  Augmented Dickey-Fuller Test
##
## data:  cumsum(x)
## Dickey-Fuller = -1.6625, Lag order = 0, p-value = 0.7212
## alternative hypothesis: stationary
```

4.3 Unit Roots

The series of length n, `x`, is a white noise process, but the vector constructed by `cumsum(x)` is a random walk.

The lag order, p, can be specified by using a `k=p+1` argument to the function.

4.3 Unit Roots

Example: S&P 500: Oct 03, 2001 - Oct 03, 2011



4.3 Unit Roots

```
sp500<-read.csv("https://www.math.mcgill.ca//dstephens//545//Data//sp500.csv")
head(sp500)

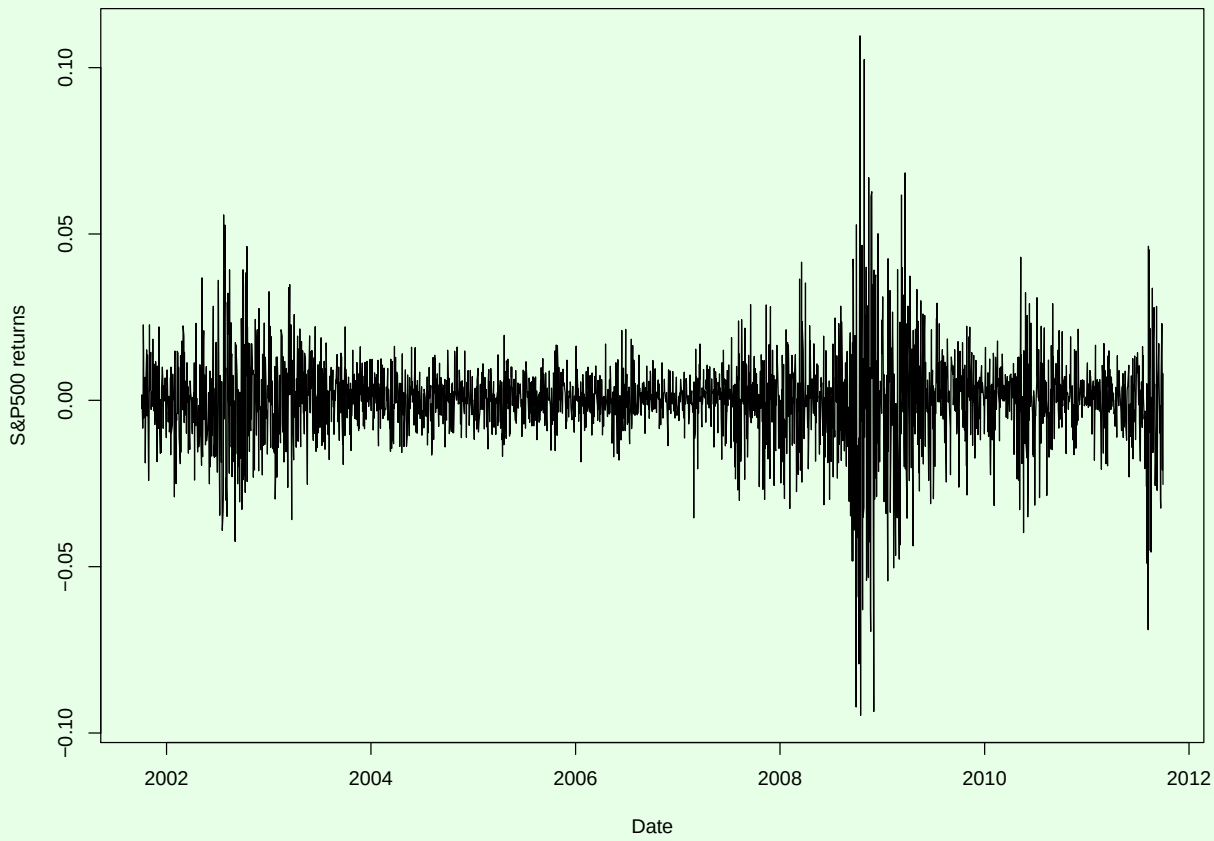
##         Date    Open    High     Low   Close     Volume Adj.Close
## 1 30/09/2011 1159.93 1159.93 1131.34 1131.42 4416790000   1131.42
## 2 29/09/2011 1151.74 1175.87 1139.93 1160.40 5285740000   1160.40
## 3 28/09/2011 1175.39 1184.71 1150.40 1151.06 4787920000   1151.06
## 4 27/09/2011 1163.32 1195.86 1163.32 1175.38 5548130000   1175.38
## 5 26/09/2011 1136.91 1164.19 1131.07 1162.95 4762830000   1162.95
## 6 23/09/2011 1128.82 1141.72 1121.36 1136.43 5639930000   1136.43

xt<-rev(sp500[,5])
adf.test(xt,k=0)

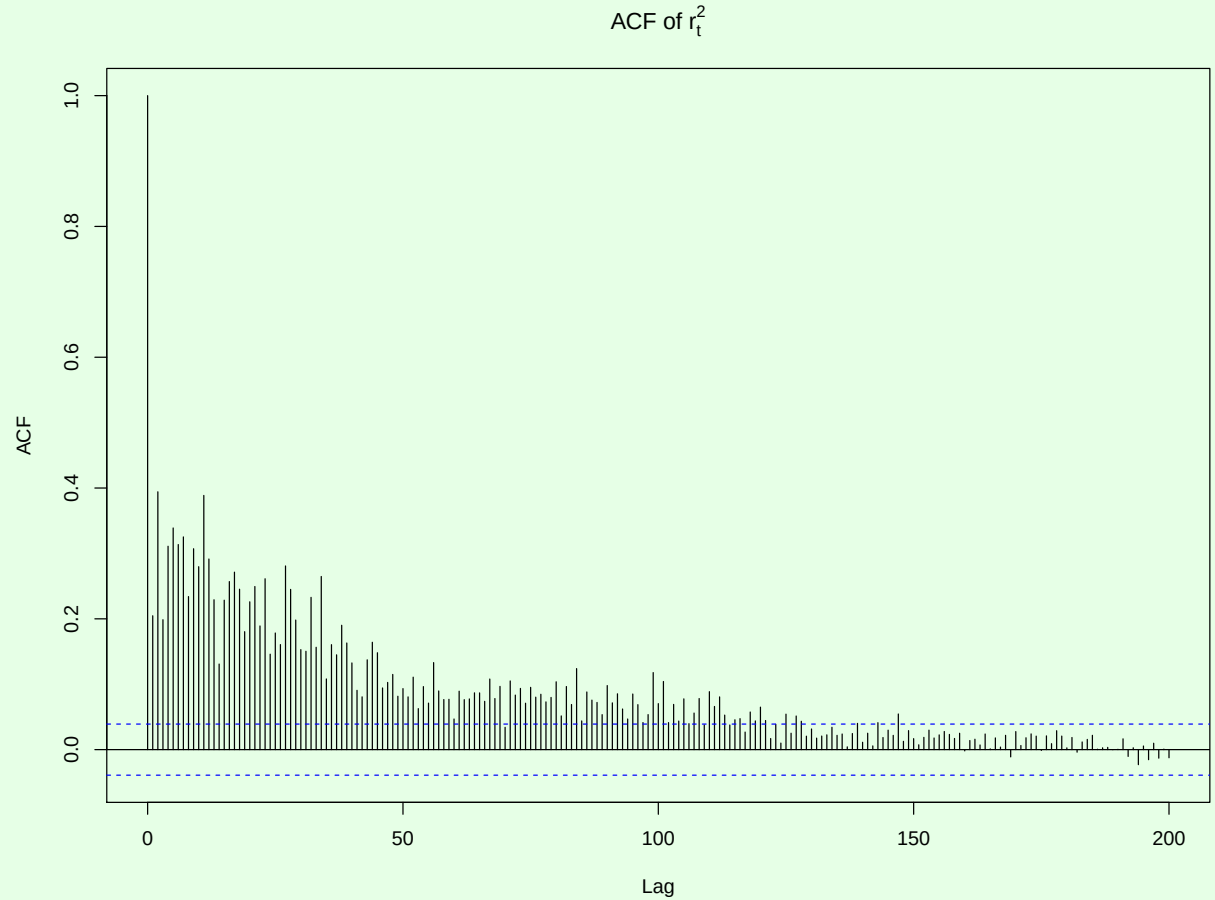
##
##  Augmented Dickey-Fuller Test
##
## data:  xt
## Dickey-Fuller = -2.0111, Lag order = 0, p-value = 0.5736
## alternative hypothesis: stationary
```

4.3 Unit Roots

Example: S&P 500: log returns $y_t = \log(x_{t+1}/x_t)$



Example: S&P 500: ACF of y_t^2



4.3 Unit Roots

Unit Roots in MA processes: If $\{X_t\} \sim \text{ARMA}(p, q)$ is causal and invertible, then

$$Y_t = (1 - B)X_t$$

is an $\text{ARMA}(p, q + 1)$ process. This follows as

$$\Phi(B)X_t = \Theta(B)Z_t$$

with if $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, then

$$\Phi(B)Y_t = \Phi(B)(1 - B)X_t = (1 - B)\Theta(B)Z_t.$$

4.3 Unit Roots

The MA polynomial for $\{Y_t\}$ is

$$\Theta_Y(z) = (1 - z)\Theta(z)$$

Thus a test for a unit root in the MA polynomial for $\{Y_t\}$ is equivalent to a test for over-differencing of a stationary, causal and invertible ARMA process.

The asymptotic null distribution of the (likelihood-ratio) test is non-standard in this setting.

4.4 Forecasting ARIMA

For the general ARIMA(p, d, q) model

$$(1 - B)^d \Phi(B) X_t = \Theta(B) Z_t$$

we perform first forecasting of

$$(1 - B)^d X_t = Y_t$$

where $\{Y_t\}$ is a causal ARMA(p, q) process. Then

$$X_t = Y_t - \sum_{j=1}^{d} \binom{d}{j} (-1)^j X_{t-j}$$

for $t > 0$.

4.4 Forecasting ARIMA

Therefore, for $h = 1, 2, \ldots,$, given $X_{1:n}$, we have that

$$\widehat{P}_n X_{n+h} = \widehat{P}_n Y_{n+h} - \sum_{j=1}^{d} \binom{d}{j} (-1)^j \widehat{P}_n X_{n+h-j}$$

for which

- $\widehat{P}_n Y_{n+h}$ is computed using the methods devised earlier for stationary processes
- $\widehat{P}_n X_{n+h-j}$ is computed recursively, noting that

  $$\widehat{P}_n X_{n+1-j} = X_{n+1-j}$$

  for $j \geqslant 1$.

4.5 Seasonal ARIMA Models

Recall the lag-s difference operator, B^s, $s \geqslant 1$,

$$(1 - B^s)X_t = X_t - X_{t-s}$$

A seasonal ARIMA model utilizes this operator: for example

$$(1 - B^s)\Phi(B)X_t = \Theta(B)Z_t$$

applies a seasonal differencing with period s.

4.5 Seasonal ARIMA Models

If d and D are non-negative integers, then $\{X_t\}$ is a seasonal ARIMA (SARIMA) process with period s

$$\{X_t\} \sim \text{ARIMA}(p, d, q) \times (P, D, Q)_s$$

of

$$Y_t = (1 - B)^d (1 - B^s)^D X_t$$

is a causal ARMA process determined by

$$\Phi(B)\Phi_S(B^s)Y_t = \Theta(B)\Theta_S(B^s)Z_t$$

where $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$.

4.5 Seasonal ARIMA Models

In this formulation

$$\begin{aligned}
\Phi(z) &= 1 - \phi_1 z - \cdots - \phi_p z^p \\
\Phi_S(z) &= 1 - \phi_1^{(S)} z - \cdots - \phi_P^{(S)} z^P \\
\Theta(z) &= 1 + \theta_1 z + \cdots + \theta_q z^q \\
\Theta_S(z) &= 1 + \theta_1^{(S)} z + \cdots + \theta_Q^{(S)} z^Q
\end{aligned}$$

Note that

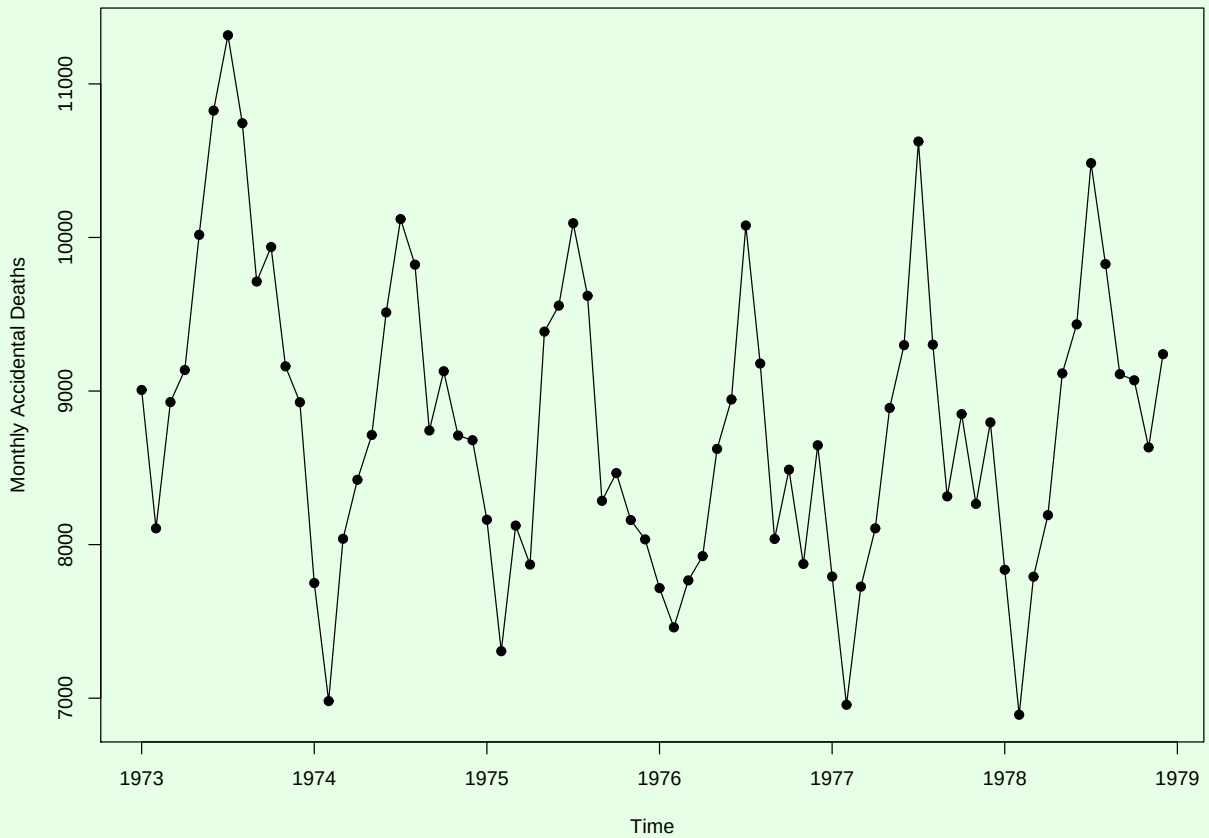
$$Y_t = (1 - B)^d (1 - B^s)^D X_t$$

is an ARMA$(p + sP, q + sQ)$ process with AR and MA polynomials

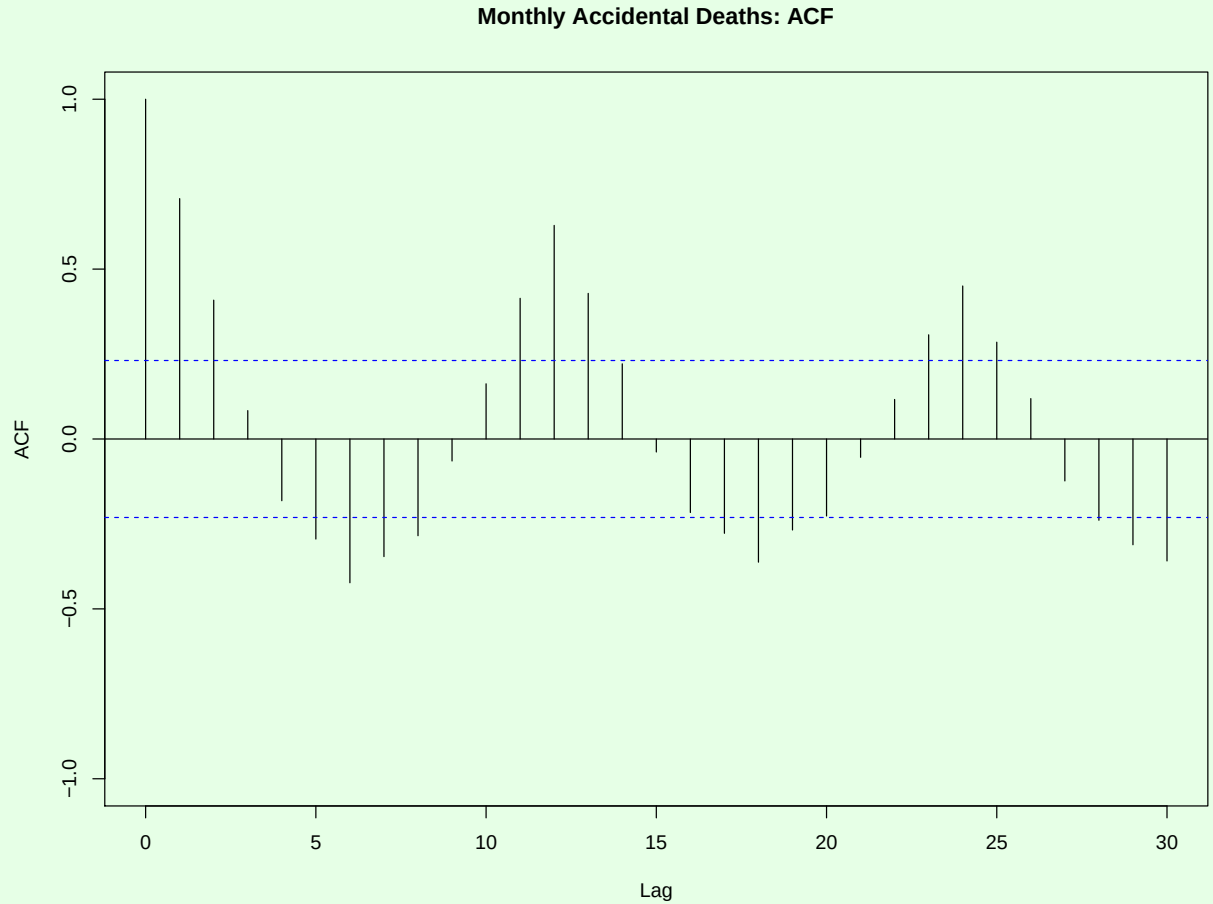
$$\Phi(B)\Phi_S(B^s) \qquad\qquad \Theta(B)\Theta_S(B^s)$$

4.5 Seasonal ARIMA Models

Example: US Accidental Deaths: Jan 1973 - Dec 1978



Example: US Accidental Deaths: Jan 1973 - Dec 1978



4.5 Seasonal ARIMA Models

Example: US Accidental Deaths: Jan 1973 - Dec 1978

Consider two SARIMA models:

A. ARIMA$(1,0,1)\times(0,1,0)_{12}$: This model fits an ARMA$(1,1)$ model to the seasonally-differenced data with $s=12$.

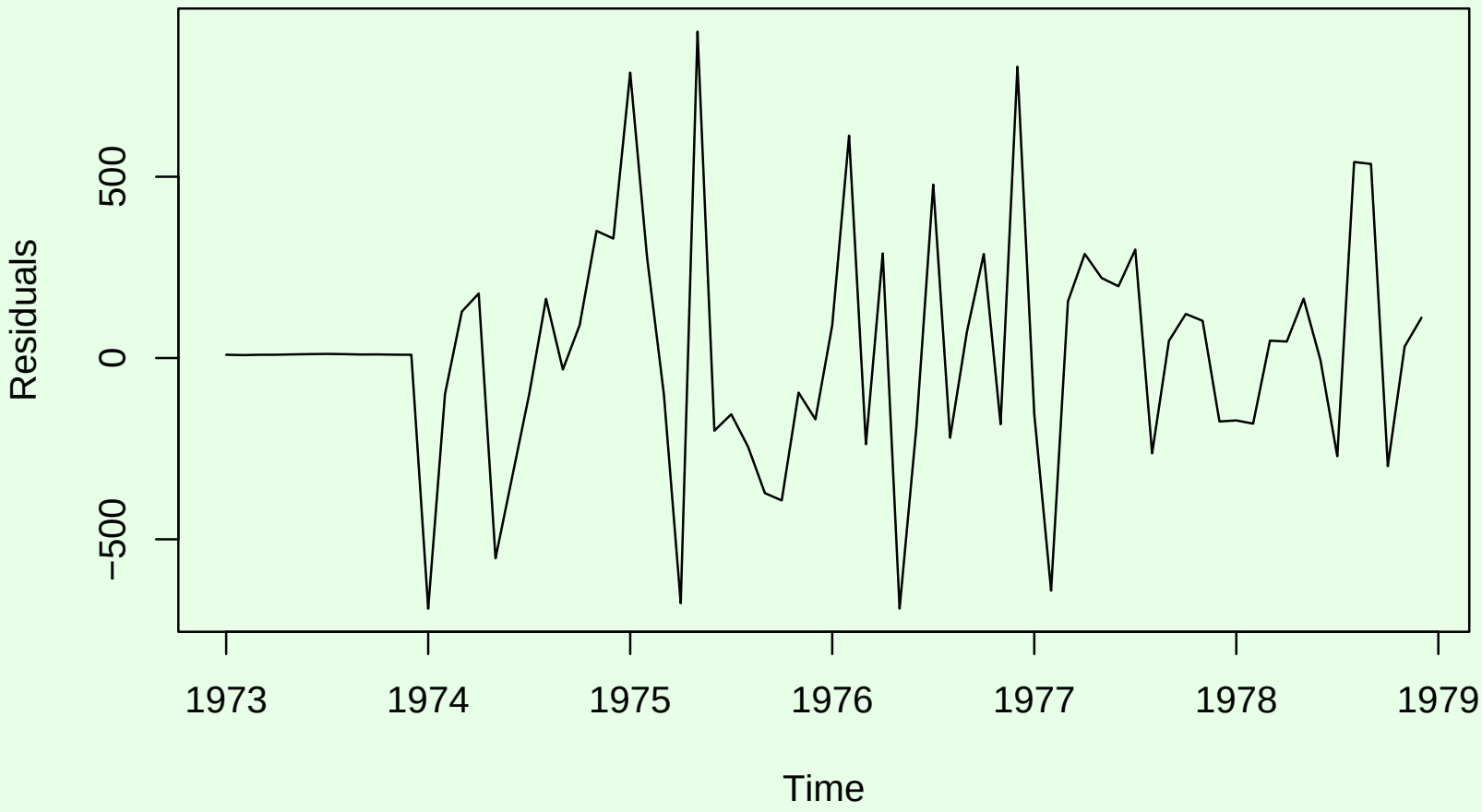
B. ARIMA$(1,0,1)\times(1,0,0)_{12}$ This model fits a SARIMA model with $p=q=1$, $P=1, Q=0$, and $s=12$.

Model A posits a non-stationary model, and requires first differencing at lag 12. Model B posits a stationary model.

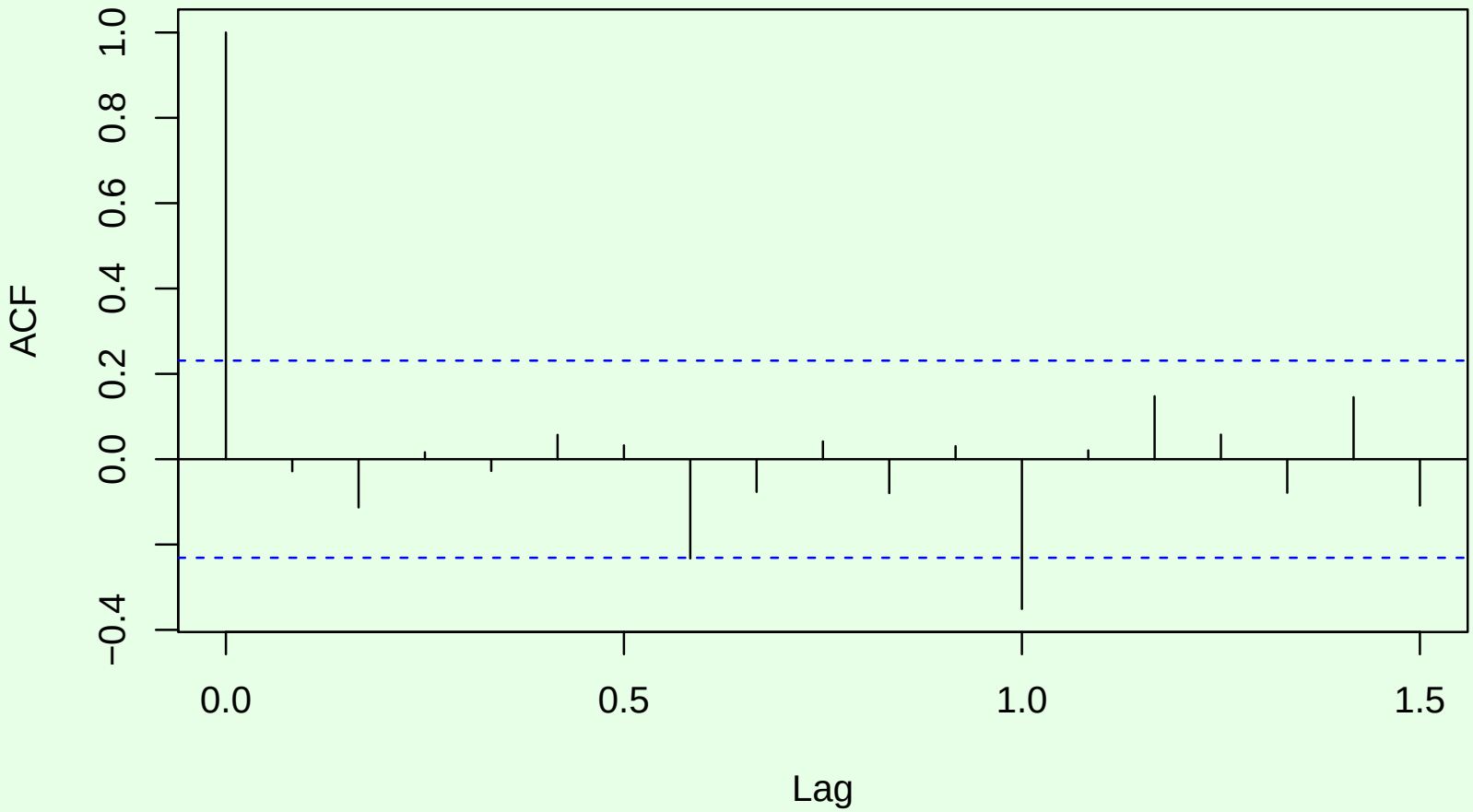
4.5 Seasonal ARIMA Models

```
model.a<-arima(USAccDeaths, order = c(1,0,1),
               seasonal = list(order=c(0,1,0),period=12))
model.b<-arima(USAccDeaths, order = c(1,0,1),
               include.mean=FALSE, seasonal = list(order=c(1,0,0),period=12))
res.a<-residuals(model.a)
res.b<-residuals(model.b)
```

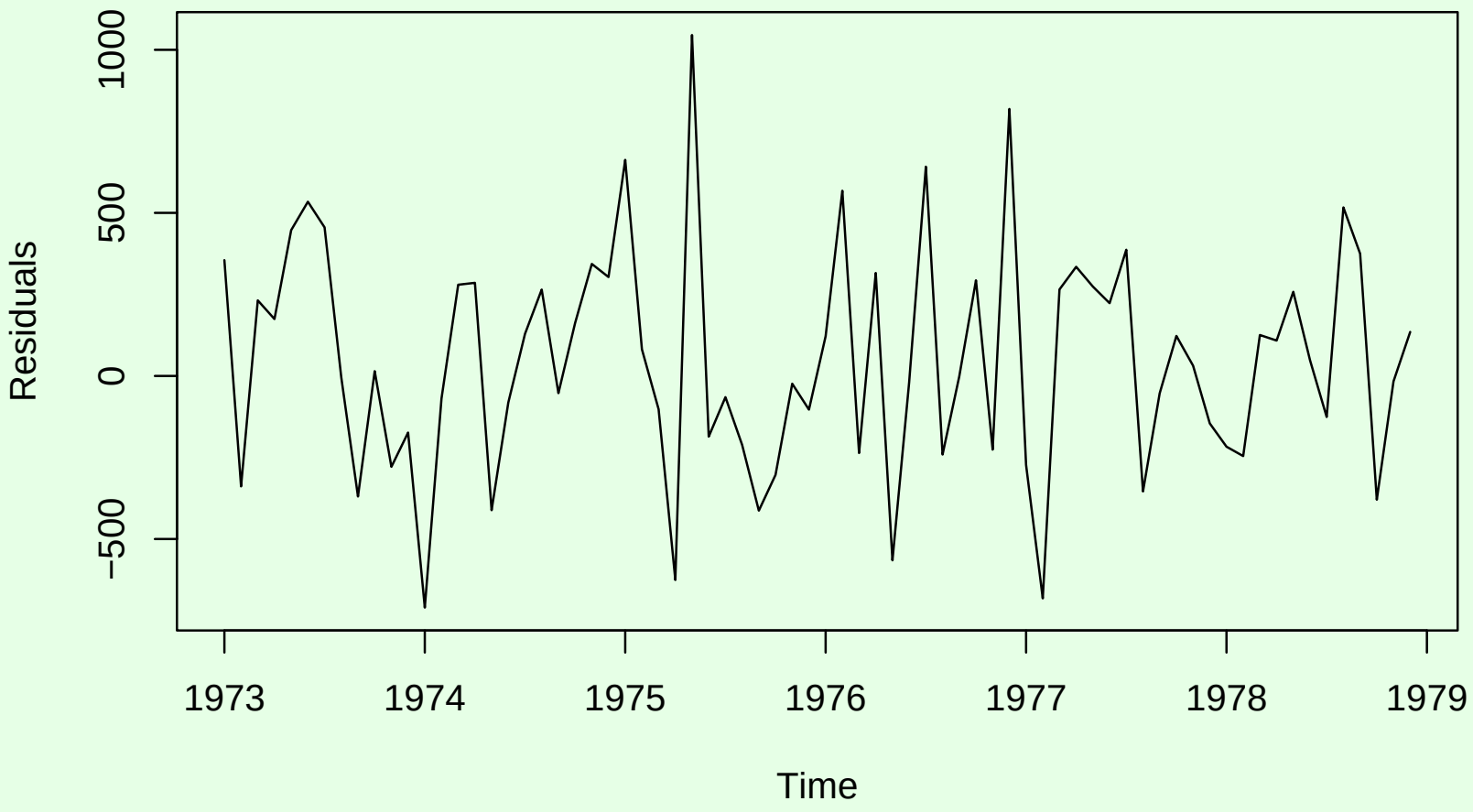
Example: Model A: Residuals



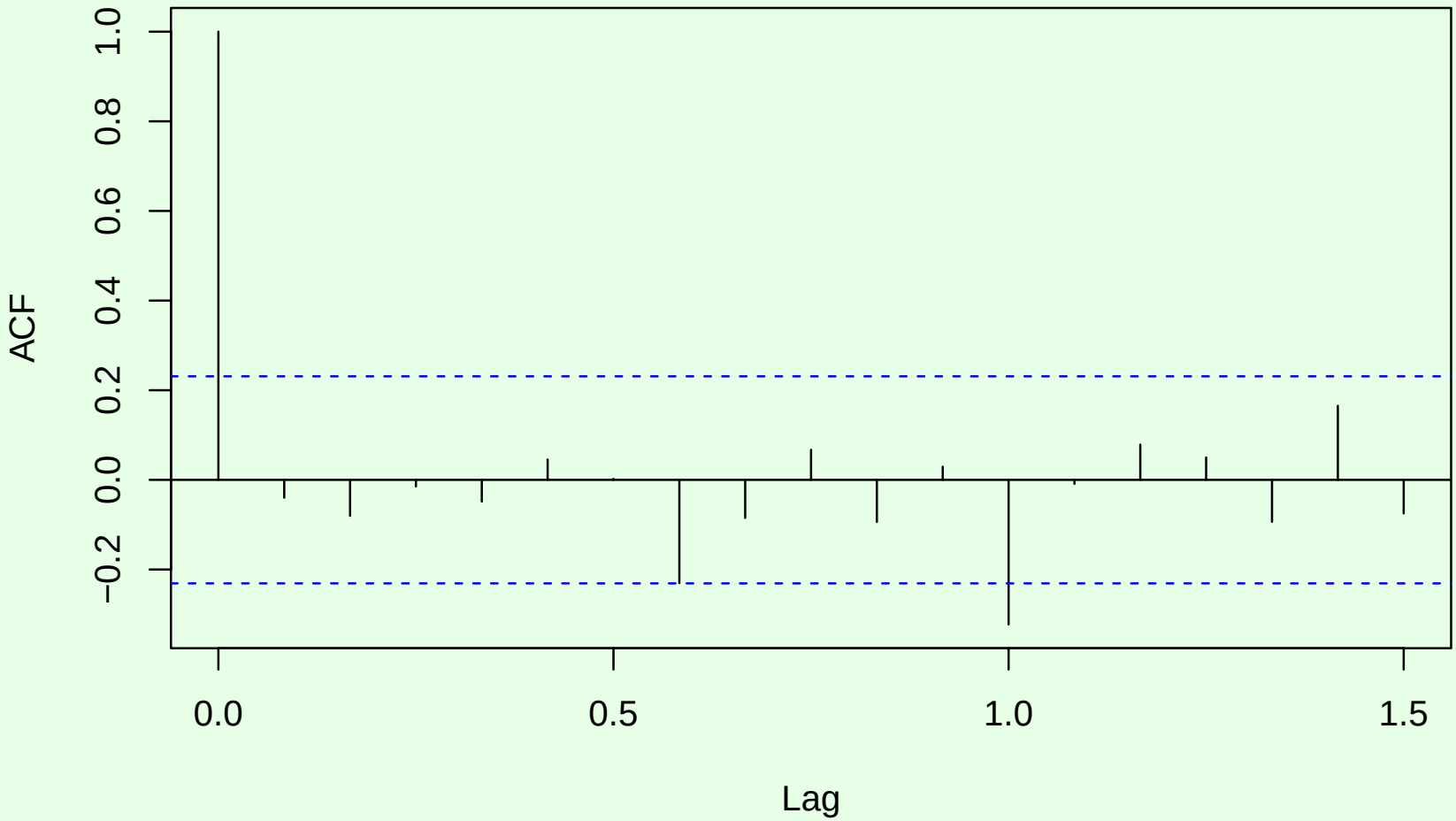
Example: Model A: Residuals ACF



Example: Model B: Residuals



Example: Model B: Residuals ACF



4.5 Seasonal ARIMA Models

Model A:

```
model.a

##
## Call:
## arima(x = USAccDeaths, order = c(1, 0, 1), seasonal = list(order = c(0, 1, 0),
##     period = 12))
##
## Coefficients:
##          ar1      ma1
##       0.9421  -0.4328
## s.e.  0.0583   0.1445
##
## sigma^2 estimated as 123848:  log likelihood = -437.61,  aic = 881.23
```

4.5 Seasonal ARIMA Models

Model B:

```
model.b

##
## Call:
## arima(x = USAccDeaths, order = c(1, 0, 1), seasonal = list(order = c(1, 0, 0),
##     period = 12), include.mean = FALSE)
##
## Coefficients:
##          ar1      ma1    sar1
##       0.9852  -0.3342  0.9098
## s.e.  0.0127   0.1368  0.0395
##
## sigma^2 estimated as 121834:  log likelihood = -537.14,  aic = 1082.29
```

4.6 Non-stationarity in Mean

Non-stationarity in mean arises when there is a time-varying mean or trend structure, as described in earlier sections. Previously OLS was used to estimate parameters in a parametric trend, m_t; OLS assumes that the residual errors are uncorrelated.

More efficient estimation can be used to take into account the correlation present if the residual error process is stationary, but exhibits autocorrelation.

4.6 Non-stationarity in Mean

Suppose

$$Y_t = m_t + X_t = \mathbf{w}_t^\top \beta + X_t$$

where

- $\mathbf{w}_t$ is a vector of (exogenous) covariates
- β is a vector of (fixed effects) coefficients
- $\{X_t\} \sim \text{ARMA}(p, q)$, with parameters (ϕ, θ)

We wish to estimate

$$\beta \qquad (\phi_1, \ldots, \phi_p)^\top \qquad (\theta_1, \ldots, \theta_q)^\top$$

4.6 Non-stationarity in Mean

Suppose

$$\text{Cov}[X_{1:n}] = \mathbb{E}[X_{1:n}X_{1:n}^\top] = \mathbf{\Gamma}_n(\phi, \theta).$$

If (ϕ, θ) are known, then we may estimate β using

Weighted Least Squares (WLS)

as

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{\Gamma}_n(\phi, \theta)^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{\Gamma}_n(\phi, \theta)^{-1}\mathbf{Y}$$

with variance

$$(\mathbf{X}^\top \mathbf{\Gamma}_n(\phi, \theta)^{-1}\mathbf{X})^{-1}$$

4.6 Non-stationarity in Mean

This suggests an iterative procedure

- Initialize $\phi^{(0)}, \theta^{(0)}$
- Compute $\beta^{(0)}$ using WLS setting $(\phi, \theta) = (\phi^{(0)}, \theta^{(0)})$.
- Compute the residuals

$$\widehat{x}_t = y_t - \mathbf{w}_t^\top \beta^{(0)}$$

  and then fit an ARMA(p, q) model to $\{\widehat{x}_t\}$ to update estimates $(\phi^{(1)}, \theta^{(1)})$.
- Continue iterations $r = 1, 2, \ldots$ until iterates at step r

$$(\beta^{(r)}, \phi^{(r)}, \theta^{(r)})$$

  cease to change.

Chapter 5
Multivariate and State Space Models

5.1 Vector Time Series

Univariate methods of time series analysis can be extended to study parallel series.

Data may comprise

- stocks in a sector, indices, foreign exchange rates
- prices of bonds of different maturities
- biomedical measurements of different features
- counts of different disease occurrences across time

all which may exhibit evolution in time in some dependent fashion.

5.2 Stationarity and Correlation

Let $\mathbf{X}_t = (X_{1,t}, \ldots, X_{K,t})^\top$ be a K-dimensional time series: $\{\mathbf{X}_t\}$ is weakly stationary if

$$\mathbb{E}[\mathbf{X}_t] = \boldsymbol{\mu} \qquad \text{Var}[\mathbf{X}_t] = \mathbb{E}[(\mathbf{X}_t - \boldsymbol{\mu})(\mathbf{X}_t - \boldsymbol{\mu})^\top] = \boldsymbol{\Gamma}_0$$

and $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}_0$ are a constant $(K \times 1)$ vector and $(K \times K)$ matrix respectively.

As usual, the diagonal elements of $\boldsymbol{\Gamma}_0$ give the variances of the K elements of the time series. They will be denoted

$$\Gamma_{11}(0), \ldots, \Gamma_{KK}(0)$$

Cross-Correlation

Let $\mathbf{D}$ be a $(K \times K)$ diagonal matrix

$$\mathbf{D} = \text{diag}(\sqrt{\Gamma_{11}(0)}, \ldots, \sqrt{\Gamma_{KK}(0)})$$

The lag-zero cross-correlation matrix $(K \times K)$ is

$$\boldsymbol{\rho}_0 = \mathbf{D}^{-1}\boldsymbol{\Gamma}_0\mathbf{D}^{-1}$$

with (i, j)th element

$$\rho_{ij}(0) = \frac{\Gamma_{ij}(0)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}}$$

This matrix gives the instantaneous correlation between the elements of the vector.

Cross-Correlation

Now consider the covariance across series at non-zero lags:

$$\mathbf{\Gamma}_h = \mathbb{E}[(\mathbf{X}_t - \boldsymbol{\mu})(\mathbf{X}_{t-h} - \boldsymbol{\mu})^\top]$$

be the lag-h cross-covariance matrix.

The (i,j)th element of this matrix is

$$\Gamma_{ij}(h) = \mathbb{E}[(X_{i,t} - \mu_i)(X_{j,t-h} - \mu_j)].$$

Cross-Correlation

Define the lag-h cross-correlation matrix as

$$\boldsymbol{\rho}_h = \mathbf{D}^{-1}\boldsymbol{\Gamma}_h\mathbf{D}^{-1}$$

with (i,j)th element

$$\rho_{ij}(h) = \frac{\Gamma_{ij}(h)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}}$$

This matrix gives the instantaneous correlation between the elements of the vector. For stationarity, these quantities can depend on h but not t.

Cross-Correlation

- $\rho_{ij}(h)$ measures the (linear) dependence of $X_{i,t}$ on $X_{j,t-h}$ which occurred prior to time t.
- $\rho_{ii}(h)$ is the lag-h autocorrelation for $\{X_{i,t}\}$.
- Note that in general

$$\rho_{ij}(h) \neq \rho_{ji}(h)$$

so $\mathbf{\Gamma}_h$ and $\boldsymbol{\rho}_h$ are not symmetric, but

$$\Gamma_{ij}(h) = \Gamma_{ji}(-h)$$

so $\mathbf{\Gamma}_h = \mathbf{\Gamma}_{-h}^\top$.

For $h > 0$, ρ_{ij} measures the dependence of $X_{i,t}$ on $X_{j,t-h}$.

Cross-Correlation: Lead-lag relationships

In summary

- if $\rho_{ij}(h) = \rho_{ji}(h) = 0$ for all $h > 0$, then $\{X_{i,t}\}$ and $\{X_{j,t}\}$ have no linear relationship.
- if $\rho_{ij}(0) \neq 0$ then $\{X_{i,t}\}$ and $\{X_{j,t}\}$ are cross-correlated at the same time point.
- if $\rho_{ij}(h) = 0$ and $\rho_{ji}(h) = 0$ for all $h > 0$, then $\{X_{i,t}\}$ and $\{X_{j,t}\}$ have no lead-lag relationship.
- if $\rho_{ij}(h) = 0$ for all $h > 0$ but $\rho_{ji}(h) \neq 0$ for some $h > 0$, then $\{X_{i,t}\}$ does not depend on past values of $\{X_{j,t}\}$, but $\{X_{j,t}\}$ does depend on past values of $\{X_{i,t}\}$.
- if $\rho_{ij}(h) \neq 0$ for some $h > 0$ and $\rho_{ji}(h) \neq 0$ for some $h > 0$, then $\{X_{i,t}\}$ and $\{X_{j,t}\}$ exhibit a feedback relationship.

Sample cross-correlation

We estimate the cross-correlation matrices by the usual moment estimators: for $h > 0$,

$$\widehat{\mathbf{\Gamma}}_h = \frac{1}{n} \sum_{t=h+1}^{n} (\mathbf{x}_t - \overline{\mathbf{x}})(\mathbf{x}_{t-h} - \overline{\mathbf{x}})^\top$$

where

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t$$

and

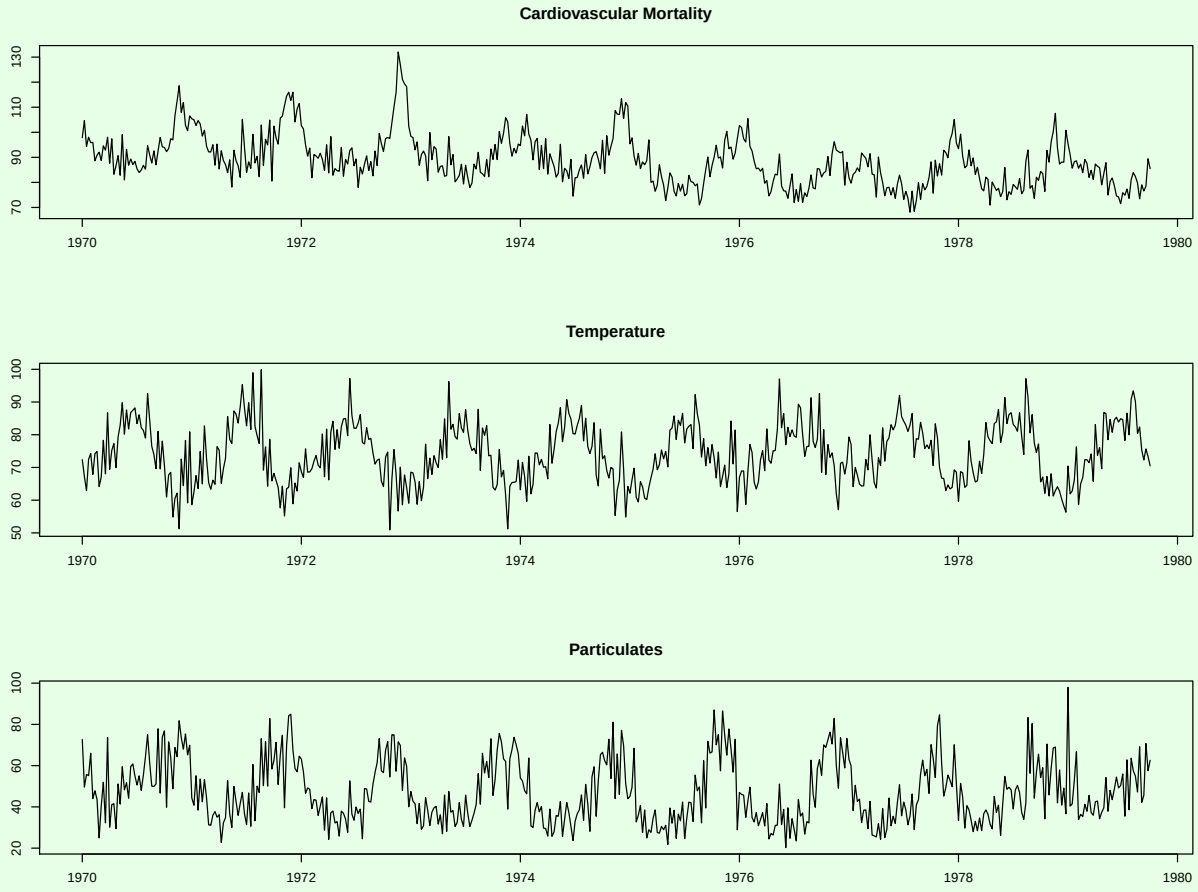
$$\widehat{\boldsymbol{\rho}}_h = \widehat{\mathbf{D}}^{-1} \widehat{\mathbf{\Gamma}}_h \widehat{\mathbf{D}}^{-1}$$

where

$$\widehat{\mathbf{D}} = \operatorname{diag}\left(\sqrt{\widehat{\Gamma}_{11}(0)}, \ldots, \sqrt{\widehat{\Gamma}_{KK}(0)} \right)$$

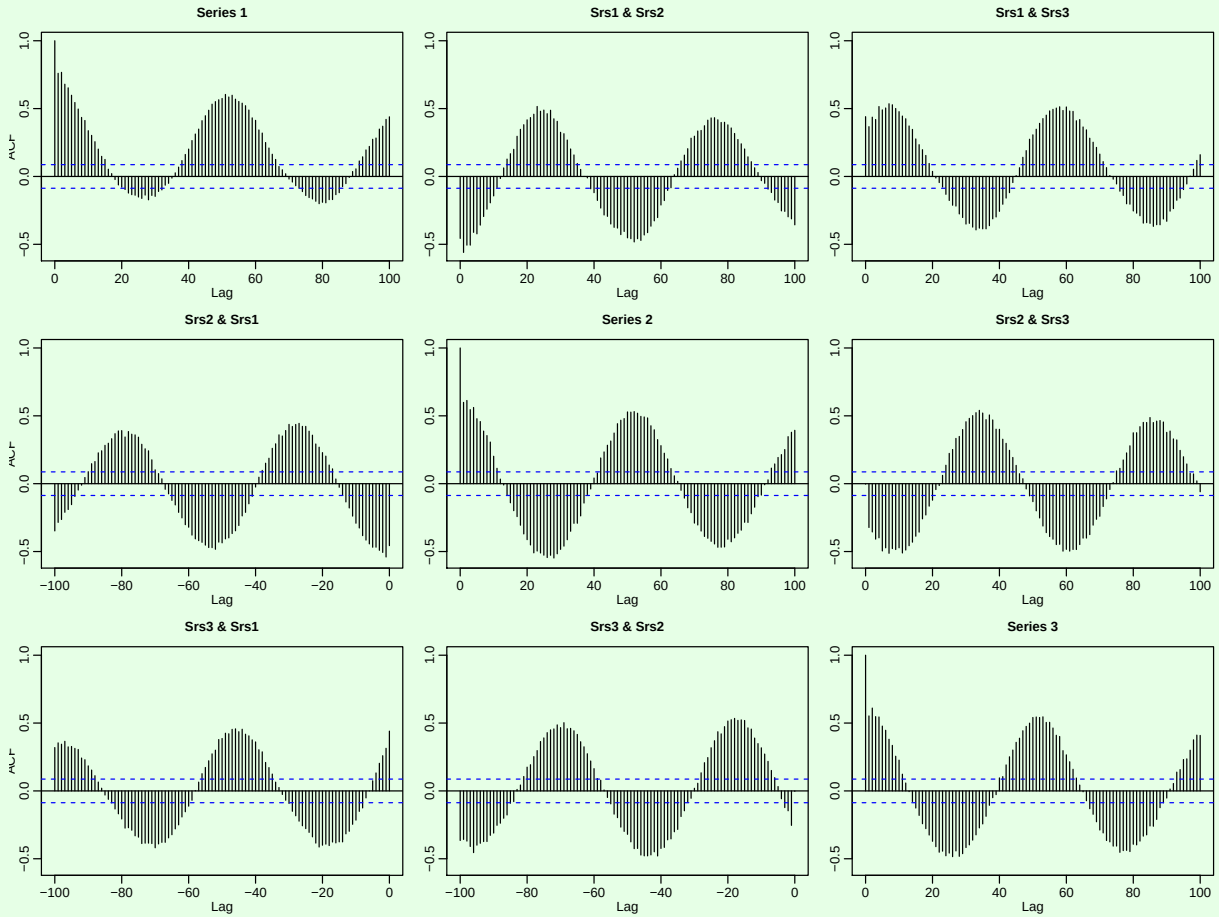
Sample cross-correlation

Example: Cardiac Mortality Data



Sample cross-correlation

Example: Cardiac Mortality Data



5.3 Vector Autoregression Models

A simple model for $\mathbf{X}_t$ takes the form

$$\mathbf{X}_t = \phi_0 + \sum_{j=1}^{p} \mathbf{\Phi}_j \mathbf{X}_{t-j} + \mathbf{Z}_t$$

where

- ϕ_0 is a $K \times 1$ vector
- $\mathbf{\Phi}_j$ is a $K \times K$ matrix
- $\mathbf{Z}_t$ is zero mean vector white noise process with

$$\mathbb{E}[\mathbf{Z}_t \mathbf{Z}_t^\top] = \Sigma$$

$\mathbf{X}_t$ follows a vector autoregression (VAR) model of order p.

5.3 Vector Autoregression Models

Usually

$$\mathbf{Z}_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

is a suitable assumption. In this case, the resulting process is a multivariate Gaussian process.

Estimation of the model can proceed using the usual likelihood methods.

Restrictions on the model are required to preserve stationarity.

Example: VAR(1)

Consider the model

$$\begin{aligned} X_{1,t} &= \phi_{10} + \phi_{11} X_{1,t-1} + \phi_{12} X_{2,t-1} + Z_{1,t} \\ X_{2,t} &= \phi_{20} + \phi_{21} X_{1,t-1} + \phi_{22} X_{2,t-1} + Z_{2,t} \end{aligned}$$

Then the elements of

$$\mathbf{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}$$

define the dependence of current elements of $\mathbf{X}_t$ on past elements.

Σ defines the concurrent relationship between $X_{1,t}$ and $X_{2,t}$.

Example: VAR(1)

Let

$$\Sigma = \mathbf{L}\mathbf{G}\mathbf{L}^\top$$

be a Cholesky decomposition where $\mathbf{G}$ is a diagonal matrix, and $\mathbf{L}$ is a lower-triangular matrix. Define $\mathbf{U}_t = \mathbf{L}^{-1}\mathbf{Z}_t$. Then

$$\mathbb{E}[\mathbf{U}_t] = \mathbf{0} \qquad \text{Var}[\mathbf{U}_t] = \mathbf{G}$$

so that the elements of $\mathbf{U}_t$ are uncorrelated. Then

$$\mathbf{L}^{-1}\mathbf{X}_t = \mathbf{L}^{-1}\phi_0 + \mathbf{L}^{-1}\mathbf{\Phi}\mathbf{X}_{t-1} + \mathbf{L}^{-1}\mathbf{Z}_t$$

so that

$$\mathbf{L}^{-1}\mathbf{X}_t = \phi^\star + \mathbf{\Phi}^\star\mathbf{X}_{t-1} + \mathbf{U}_t$$

say.

Example: VAR(1)

Thus, as $\mathbf{L}$ is lower-triangular, the Kth row of this vector equation is

$$X_{K,t} + \sum_{k=1}^{K-1} L_{Kk} X_{k,t} = \phi^{\star}_{K0} + \sum_{k=1}^{K} \phi^{\star}_{Kk} X_{k,t-1} + U_{K,t}$$

As the elements of $\mathbf{U}_t$ are uncorrelated, this equation shows explicitly the concurrent dependence of $X_{K,t}$ on $X_{k,t}$, for $k = 1, \ldots, K-1$.

This is termed the structural equation for $X_{K,t}$. A similar expression can be found for each of the elements of $\mathbf{X}_t$.

Example: $K = 2$

Let

$$\begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix} \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \end{bmatrix} + \begin{bmatrix} Z_{1,t} \\ Z_{2,t} \end{bmatrix}$$

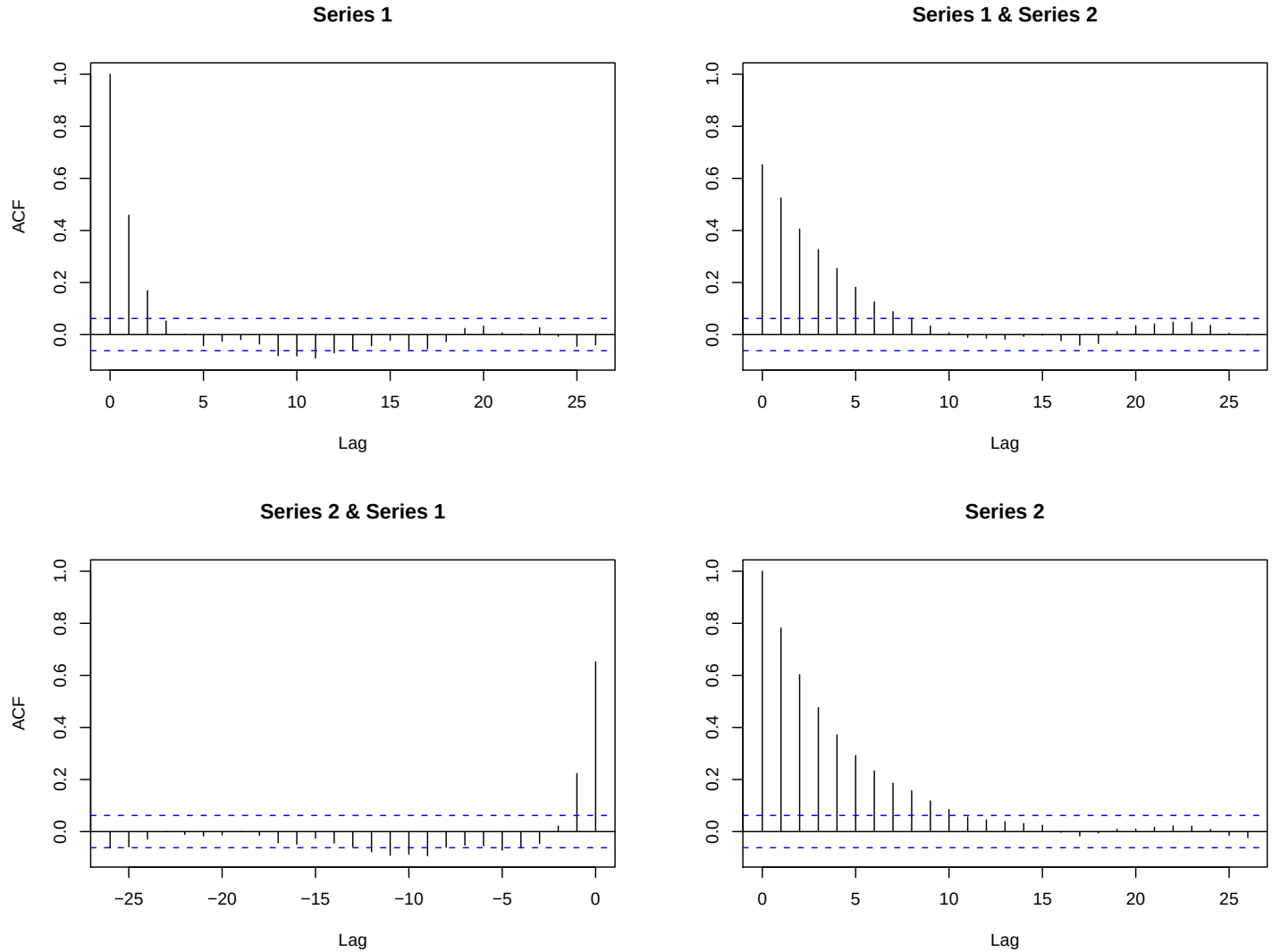
with

$$\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

Then

$$\mathbf{L}^{-1} = \begin{bmatrix} 1.0 & 0.0 \\ -0.5 & 1.0 \end{bmatrix} \qquad \mathbf{G} = \mathbf{L}^{-1}\Sigma(\mathbf{L}^{\top})^{-1} = \begin{bmatrix} 2.0 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}$$

Example: $K = 2$ - cross-correlations



Example: VAR(1)

Using the above structural equation formulation:

$$X_{2,t} = 0.3 + 0.5X_{1,t} - 0.7X_{1,t-1} + 0.95X_{2,t-1} + U_{2,t}$$

and

$$X_{1,t} = -0.2 + 1.0X_{2,t} - 0.8X_{2,t-1} + 0.8X_{1,t-1} + U_{1,t}$$

where $U_{2,t} \sim \mathcal{N}(0, 0.5)$ and $U_{1,t} \sim \mathcal{N}(0, 1)$ are independent random variables.

These equations demonstrate the concurrent linear dependence between $X_{1,t}$ and $X_{2,t}$.

Stationarity for VAR(1)

It follows from

$$\mathbf{X}_t = \phi_0 + \mathbf{\Phi}_1 \mathbf{X}_{t-1} + \mathbf{Z}_t$$

that, taking expectations, from stationarity

$$\boldsymbol{\mu} = \phi_0 + \mathbf{\Phi}_1 \boldsymbol{\mu} \qquad \therefore \qquad \boldsymbol{\mu} = (\mathbf{I}_K - \mathbf{\Phi}_1)^{-1} \phi_0$$

provided $(\mathbf{I}_K - \mathbf{\Phi}_1)$ is non-singular. Then

$$\mathbf{X}_t - \boldsymbol{\mu} = \mathbf{\Phi}_1 (\mathbf{X}_{t-1} - \boldsymbol{\mu}) + \mathbf{Z}_t$$

If $\mathbf{Y}_t = \mathbf{X}_t - \boldsymbol{\mu}$, then by recursion

$$\mathbf{Y}_t = \mathbf{Z}_t + \mathbf{\Phi}_1 \mathbf{Z}_{t-1} + \mathbf{\Phi}_1^2 \mathbf{Z}_{t-2} + \cdots$$

and hence $\text{Cov}[\mathbf{Z}_t, \mathbf{Y}_{t-h}] = \text{Cov}[\mathbf{Z}_t, \mathbf{X}_{t-h}] = \mathbf{0}$, for all $h > 0$.

Stationarity for VAR(1)

In addition,

$$\mathbf{Y}_t\mathbf{Z}_t^\top = \mathbf{Z}_t\mathbf{Z}_t^\top + \mathbf{\Phi}_1\mathbf{Z}_{t-1}\mathbf{Z}_t^\top + \mathbf{\Phi}_1^2\mathbf{Z}_{t-2}\mathbf{Z}_t^\top + \cdots$$

and taking expectations yields

$$\text{Cov}[\mathbf{Y}_t, \mathbf{Z}_t] = \mathbb{E}[\mathbf{Y}_t\mathbf{Z}_t^\top] = \mathbb{E}[\mathbf{Z}_t\mathbf{Z}_t^\top] = \Sigma$$

as $\{\mathbf{Z}_t\}$ is a white-noise process.

Stationarity for VAR(1)

In the representation, $\mathbf{X}_t$ depends on $\mathbf{Z}_{t-j}$ with coefficient matrix $\mathbf{\Phi}_1^j$; this dependence must converge to zero. Therefore, using the eigendecomposition

$$\mathbf{\Phi}_1 = \mathbf{A}_1 \mathbf{\Lambda}_1 \mathbf{B}_1$$

with $\mathbf{A}_1 \mathbf{B}_1 = \mathbf{I}_K$ and $\mathbf{\Lambda}_1 = \text{diag}(\lambda_{11}, \ldots, \lambda_{1K})$, we have in the usual way

$$\mathbf{\Phi}_1^j = \mathbf{A}_1 \mathbf{\Lambda}_1^j \mathbf{B}_1.$$

Stationarity for VAR(1)

Thus, for each j,

$$\mathbf{\Phi}_1^j \longrightarrow \mathbf{0} \qquad \Longleftrightarrow \qquad |\lambda_{1k}| < 1,\ k = 1, \ldots, K$$

as in the univariate case. Note that

$$\det(\lambda \mathbf{I}_K - \mathbf{\Phi}_1) = \lambda^K \det(\mathbf{I}_K - \lambda^{-1} \mathbf{\Phi}_1)$$

and hence $\lambda_{11}, \ldots, \lambda_{1K}$ are found via the solutions of

$$\det(\mathbf{I}_K - \mathbf{\Phi}_1 z) = 0.$$

Stationarity for VAR(1)

Similar to the univariate case, $\mathbf{X}_t$ is stationary if and only if the roots of

$$\det(z\mathbf{I}_K - \mathbf{\Phi}_1) = 0.$$

are greater than 1 in modulus.

Finally

$$\mathbf{\Gamma}_0 = \text{Var}[\mathbf{X}_t] = \text{Var}[\mathbf{Z}_t] + \sum_{j=1}^{\infty} \text{Var}[\mathbf{\Phi}_1^j \mathbf{Z}_{t-j}] = \Sigma + \sum_{j=1}^{\infty} \mathbf{\Phi}_1^j \Sigma (\mathbf{\Phi}_1^\top)^j$$

Stationarity for VAR(1)

Using the equation

$$\mathbf{X}_t - \boldsymbol{\mu} = \boldsymbol{\Phi}_1(\mathbf{X}_{t-1} - \boldsymbol{\mu}) + \mathbf{Z}_t$$

we have for $h > 0$,

$$\begin{aligned}\mathbb{E}[(\mathbf{X}_t - \boldsymbol{\mu})(\mathbf{X}_{t-h} - \boldsymbol{\mu})^\top] \\ = \boldsymbol{\Phi}_1\mathbb{E}[(\mathbf{X}_{t-1} - \boldsymbol{\mu})(\mathbf{X}_{t-h} - \boldsymbol{\mu})^\top] + \mathbb{E}[\mathbf{Z}_t(\mathbf{X}_{t-h} - \boldsymbol{\mu})^\top]\end{aligned}$$

or equivalently, as the final expectation is zero

$$\boldsymbol{\Gamma}_h = \boldsymbol{\Phi}_1\boldsymbol{\Gamma}_{h-1} = \boldsymbol{\Phi}_1^h\boldsymbol{\Gamma}_0$$

by recursion.

Stationarity for VAR(1)

Thus

$$\begin{aligned}
\boldsymbol{\rho}_h &= \mathbf{D}^{-1}\boldsymbol{\Phi}_1\boldsymbol{\Gamma}_{h-1}\mathbf{D}^{-1} \\
&= \left\{\mathbf{D}^{-1}\boldsymbol{\Phi}_1\mathbf{D}\right\}\left\{\mathbf{D}^{-1}\boldsymbol{\Gamma}_{h-1}\mathbf{D}^{-1}\right\} \\
&= \boldsymbol{\Upsilon}_1\boldsymbol{\rho}_{h-1} \\
&\vdots \\
&= \boldsymbol{\Upsilon}_1^h\boldsymbol{\rho}_0
\end{aligned}$$

Note: $\mathbf{D} = \text{diag}(\sqrt{\Gamma_{11}(0)}, \ldots, \sqrt{\Gamma_{KK}(0)})$ so square root already taken.

Results for VAR(p) models

We have

$$\mathbf{X}_t - \boldsymbol{\mu} = \sum_{j=1}^{p} \boldsymbol{\Phi}_j (\mathbf{X}_{t-j} - \boldsymbol{\mu}) + \mathbf{Z}_t$$

and

$$\text{Cov}[\mathbf{X}_t, \mathbf{Z}_t] = \Sigma \qquad \text{Cov}[\mathbf{X}_{t-h}, \mathbf{Z}_t] = \mathbf{0}, \ h > 0$$

Also, for $h > 0$,

$$\boldsymbol{\Gamma}_h = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{\Gamma}_{h-j} \qquad \boldsymbol{\rho}_h = \sum_{j=1}^{p} \boldsymbol{\Upsilon}_j \boldsymbol{\rho}_{h-j}$$

where $\boldsymbol{\Upsilon}_j = \mathbf{D}^{-1/2} \boldsymbol{\Phi}_j \mathbf{D}^{-1/2}$

Results for VAR(p) models

As for the VAR(1),

$$\mathbf{X}_t = \boldsymbol{\mu} + \mathbf{Z}_t + \boldsymbol{\Psi}_1 \mathbf{Z}_{t-1} + \boldsymbol{\Psi}_2 \mathbf{Z}_{t-2} + \cdots$$

by the usual moving average type construction, where

$$(\mathbf{I}_K - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p)(\mathbf{I}_K + \boldsymbol{\Psi}_1 B + \boldsymbol{\Psi}_2 B^2 + \cdots) = \mathbf{I}_K$$

The matrix $\boldsymbol{\Psi}_j$ is the effect of $\mathbf{Z}_t$ on $\mathbf{X}_{t+j}$; it is termed the impulse response function of $\mathbf{X}_t$.

The same trick to orthogonalize the system by writing

$$\Sigma = \mathbf{L}\mathbf{G}\mathbf{L}^\top$$

and pre-multiplying by $\mathbf{L}^{-1}$ can be used.

Vector MA Models

Let

$$\mathbf{X}_t = \boldsymbol{\theta}_0 + \mathbf{Z}_t + \boldsymbol{\Theta}_1 \mathbf{Z}_{t-1} + \cdots + \boldsymbol{\Theta}_q \mathbf{Z}_{t-q} = \boldsymbol{\theta}_0 + \boldsymbol{\Theta}(B)\mathbf{Z}_t$$

where the $\boldsymbol{\Theta}_j$ are $K \times K$ matrices, and $\boldsymbol{\theta}_0$ is $K \times 1$. Then

- $\text{Cov}[\mathbf{X}_t, \mathbf{Z}_t] = \Sigma$
- $\boldsymbol{\Gamma}_0 = \Sigma + \boldsymbol{\Theta}_1 \Sigma \boldsymbol{\Theta}_1^\top + \cdots + \boldsymbol{\Theta}_q \Sigma \boldsymbol{\Theta}_q^\top$
- $$\boldsymbol{\Gamma}_h = \begin{cases} \sum_{j=1}^{q} \boldsymbol{\Theta}_j \Sigma \boldsymbol{\Theta}_{j-h} & 1 \leqslant h \leqslant q \\ 0 & h > q \end{cases}$$

  where $\boldsymbol{\Theta}_0 = \mathbf{I}_K$.

Vector ARMA Models

Let

$$\mathbf{\Phi}(B)\mathbf{X}_t = \phi_0 + \mathbf{\Theta}(B)\mathbf{Z}_t$$

where

$$\begin{aligned}\mathbf{\Phi}(B) &= \mathbf{I}_K - \mathbf{\Phi}_1 B - \cdots - \mathbf{\Phi}_p B^p \\ \mathbf{\Theta}(B) &= \mathbf{I}_K + \mathbf{\Theta}_1 B + \cdots + \mathbf{\Theta}_q B^q\end{aligned}$$

This describes a vector ARMA(p, q) (VARMA(p, q)) model.

The marginal processes are ARMA$(Kp, (K-1)p + q)$ processes.

5.4 Cointegration

Consider the VARMA$(1,1)$ model

$$\mathbf{\Phi}(B)\mathbf{X}_t = \mathbf{\Theta}(B)\mathbf{Z}_t$$

with

$$\begin{aligned}\mathbf{\Phi}(B) &= \mathbf{\Phi}_1(B) = \begin{bmatrix} 1-0.5B & B \\ 0.25B & 1-0.5B \end{bmatrix} \\ \mathbf{\Theta}(B) &= \mathbf{\Theta}_1(B) = \begin{bmatrix} 1-0.2B & 0.4B \\ 0.1B & 1-0.2B \end{bmatrix}\end{aligned}$$

with

$$\text{Var}[\mathbf{Z}_t] = \Sigma$$

5.4 Cointegration

That is

$$\begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix} = \begin{bmatrix} 0.5 & -1 \\ -0.25 & 0.5 \end{bmatrix} \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \end{bmatrix}$$

$$+ \begin{bmatrix} Z_{1,t} \\ Z_{2,t} \end{bmatrix} + \begin{bmatrix} -0.2 & 0.4 \\ 0.1 & -0.2 \end{bmatrix} \begin{bmatrix} Z_{1,t-1} \\ Z_{2,t-1} \end{bmatrix}$$

or

$$\mathbf{X}_t = \mathbf{\Phi} \mathbf{X}_{t-1} + \mathbf{Z}_t + \mathbf{\Theta} \mathbf{Z}_{t-1}.$$

say.

5.4 Cointegration

Then

$$\begin{bmatrix} 1-0.5B & B \\ 0.25B & 1-0.5B \end{bmatrix} \mathbf{X}_t = \begin{bmatrix} 1-0.2B & 0.4B \\ 0.1B & 1-0.2B \end{bmatrix} \mathbf{Z}_t$$

Filtering on both sides by

$$\begin{bmatrix} 1-0.5B & -B \\ -0.25B & 1-0.5B \end{bmatrix}$$

we obtain on the left and right hand sides respectively

$$\begin{bmatrix} 1-0.5B & -B \\ -0.25B & 1-0.5B \end{bmatrix} \begin{bmatrix} 1-0.5B & B \\ 0.25B & 1-0.5B \end{bmatrix} = \begin{bmatrix} 1-B & 1 \\ 0 & 1-B \end{bmatrix}$$

$$\begin{bmatrix} 1-0.5B & -B \\ -0.25B & 1-0.5B \end{bmatrix} \begin{bmatrix} 1-0.2B & 0.4B \\ 0.1B & 1-0.2B \end{bmatrix} = \begin{bmatrix} 1-0.7B & -0.6B \\ -0.15B & 1-0.7B \end{bmatrix}.$$

5.4 Cointegration

Therefore

$$\begin{bmatrix} 1-B & 0 \\ 0 & 1-B \end{bmatrix} \mathbf{X}_t = \begin{bmatrix} 1-0.7B & -0.6B \\ -0.15B & 1-0.7B \end{bmatrix} \mathbf{Z}_t$$

That is

$$\begin{bmatrix} (1-B)X_{1,t} \\ (1-B)X_{2,t} \end{bmatrix} = \begin{bmatrix} Z_{1,t} - 0.7Z_{1,t-1} - 0.6Z_{2,t-1} \\ Z_{2,t} - 0.15Z_{1,t-1} - 0.7Z_{2,t-1} \end{bmatrix}$$

Procesess $X_{1,t}$ and $X_{2,t}$ have unit roots, specifically

$$X_{1,t}, X_{2,t} \sim \text{ARIMA}(0,1,1).$$

5.4 Cointegration

However, let

$$\mathbf{L} = \begin{bmatrix} 1 & -2 \\ 0.5 & 1 \end{bmatrix}$$

and let

$$\mathbf{Y}_t = \mathbf{L}\mathbf{X}_t \qquad \mathbf{U}_t = \mathbf{L}\mathbf{Z}_t$$

Then

$$\begin{aligned} \mathbf{Y}_t = \mathbf{L}\mathbf{X}_t &= \mathbf{L}\Phi\mathbf{X}_{t-1} + \mathbf{L}\mathbf{Z}_t + \mathbf{L}\Theta\mathbf{Z}_{t-1} \\ &= \mathbf{L}\Phi\mathbf{L}^{-1}\left(\mathbf{L}\mathbf{X}_{t-1}\right) + \mathbf{L}\mathbf{Z}_t + \mathbf{L}\Theta\mathbf{L}^{-1}\left(\mathbf{L}\mathbf{Z}_{t-1}\right) \\ &= \mathbf{L}\Phi\mathbf{L}^{-1}\mathbf{Y}_{t-1} + \mathbf{U}_t + \mathbf{L}\Theta\mathbf{L}^{-1}\mathbf{U}_{t-1} \end{aligned}$$

or equivalently

$$\mathbf{Y}_t - \Phi^\star\mathbf{Y}_{t-1} = \mathbf{U}_t + \Theta^\star\mathbf{U}_{t-1}$$

In this equation

$$\mathbf{\Phi}^\star = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{\Theta}^\star = \begin{bmatrix} -0.4 & 0 \\ 0 & 0 \end{bmatrix}$$

so that

$$\begin{aligned} Y_{1,t} - Y_{1,t-1} &= U_{1,t} - 0.4U_{1,t-1} \\ Y_{2,t} &= U_{2,t} \end{aligned}$$

Hence $\{Y_{1t}\} \sim \text{ARIMA}(0,1,1)$ and $\{Y_{2t}\}$ is a white noise process, and there is no dependence between the two series.

Hence, the unit roots in $\{X_{1,t}\}$ and $\{X_{2,t}\}$ result from the unit root in $\{Y_{1,t}\}$ and a subsequent linear transformation.

5.4 Cointegration

This phenomenon is known as cointegration.

Cointegration can be used to analyze co-movements in raw series (asset prices, exchange rates or yields).

The modelling strategy allows the detection of stable or stationary long-run relationships between non-stationary variables; it allows the underlying relationships between the series to be discovered.

5.4 Cointegration

The components of $K-$vector $\mathbf{X}_t$ are said to be cointegrated of order (d, b), denoted

$$\mathbf{X}_t \sim CI(d, b)$$

if

(a) the components of $\mathbf{X}_t$ are $I(d)$ (stationary after d-times differencing),

5.4 Cointegration

(b) there exists a linear combination of $\mathbf{X}_t$, $\widetilde{X}_t$ say, where

$$\widetilde{X}_t = \boldsymbol{\alpha}^\top \mathbf{X}_t \qquad \boldsymbol{\alpha} \neq \mathbf{0}$$

such that $\widetilde{X}_t \sim I(d-b)$ for some $b > 0$.

The vector $\boldsymbol{\alpha}$ is termed the cointegrating vector.

Time series that cointegrate indicate that stable long-run relationships (equilibria) between the series exist, irrespective of the behaviour of the individual series.

Estimation of Cointegrating Series

Consider $\{Y_t\}, \{\mathbf{X}_t\} = \{(X_{1,t}, \ldots, X_{K,t})^\top\} \sim I(1)$ that are cointegrated. Suppose that

$$Y_t - (\alpha_1 X_{1,t} + \ldots + \alpha_K X_{K,t}) \sim I(0)$$

that is, this process of 'residuals' is stationary.

We seek a statistical procedure to analyze this system.

Estimation of Cointegrating Series

1. Fit the regression model

$$Y_t = \alpha_1 X_{1,t} + \ldots + \alpha_K X_{K,t} + Z_t^\star$$

with (cointegrating) vector

$$\boldsymbol{\alpha} = (1, -\alpha_1, \ldots, -\alpha_K)^\top$$

and estimate the parameters in the model (using OLS or maximum likelihood).

It can be shown that the estimators are consistent with convergence rate $\mathrm{o}_p(n^{-1})$.

Estimation of Cointegrating Series

2. Under cointegration, residuals $\widehat{z}_t^{\star}$ should be $I(0)$ (stationary); this can be tested using Dickey-Fuller procedures.

   If $\widehat{z}_t^{\star} \sim I(0)$ is acceptable (that is, the unit root hypothesis in the Dickey-Fuller test is rejected) then a second regression model known as an Error Correction Model (ECM) can be specified.

Estimation of Cointegrating Series

For $K = 1$: Suppose $X_t, Y_t \sim I(1)$, and we conclude from the previous analysis that $\widehat{z}_t^{\star} \sim I(0)$. Let

$$X_t^* = (1 - B)X_t \sim I(0)$$

$$Y_t^* = (1 - B)Y_t \sim I(0)$$

be the differenced series, which are stationary.

Estimation of Cointegrating Series

Consider the regression models

$$Y_t^* = \psi_0 + \gamma_1 \widehat{z}_{t-1}^\star + \sum_{j=1}^{p} \psi_{1j} X_{t-j}^* + \sum_{l=1}^{m} \psi_{2l} Y_{t-l}^* + \epsilon_{1,t}$$

$$X_t^* = \xi_0 + \gamma_2 \widehat{z}_{t-1}^\star + \sum_{j=1}^{p} \xi_{1j} Y_{t-j}^* + \sum_{l=1}^{m} \xi_{2l} X_{t-l}^* + \epsilon_{2,t}$$

with estimated using standard OLS/ML techniques.

Estimation of Cointegrating Series

The ECM in the first equation states that changes in the series Y_t are explained by

- the residual in the long-run equilibrium from previous time point (coefficient γ_1),
- lagged changes in the differenced X_t series, via coefficients ψ_1),
- their own history via coefficients ψ_2.

Estimation of Cointegrating Series

Furthermore

- γ_1 determines the rate of re-adjustment; should have

$$\widehat{\gamma}_1 < 0.$$

- Model orders p and m need to be selected; typically start with p, m large, and drop variables according to hypothesis tests or using model selection criteria
- Can extend to systems of cointegrated variables

Two cointegrating series

Example: Cointegrating series

Suppose that

$$X_t \sim I(1)$$

$$Y_t = 0.6X_t + Z_{2,t}$$

where $Z_{2,t} \sim \mathcal{N}(0,1)$ is an IID series. This implies that

$$Y_t - 0.6X_t = Z_{2,t}$$

is a cointegrating system with $\boldsymbol{\alpha} = (1, -0.6)^\top$.

Two cointegrating series

Example: Cointegrating series

To estimate the system parameters, we regress Y_t on X_t, and then form the error correction models.

Regressions are based on the differenced series

$$Y_t^* = \psi_0 + \gamma_1 \widehat{z}_{t-1}^\star + \psi_{11} X_{t-1}^* + \psi_{21} Y_{t-1}^* + \epsilon_{1,t}$$

$$X_t^* = \xi_0 + \gamma_2 \widehat{z}_{t-1}^\star + \xi_{11} Y_{t-1}^* + \xi_{21} X_{t-1}^* + \epsilon_{2,t}$$

where $\{\widehat{z}_t^\star\}$ are the residuals from the regression of Y_t on X_t.

Simulated Example: R code

Suppose

$$X_t = X_{t-1} + Z_{1,t}$$
$$Y_t = 0.6X_t + Z_{2,t}$$

where $\{Z_{1,t}\}, \{Z_{2,t}\} \sim \text{IID}(0, 1)$.

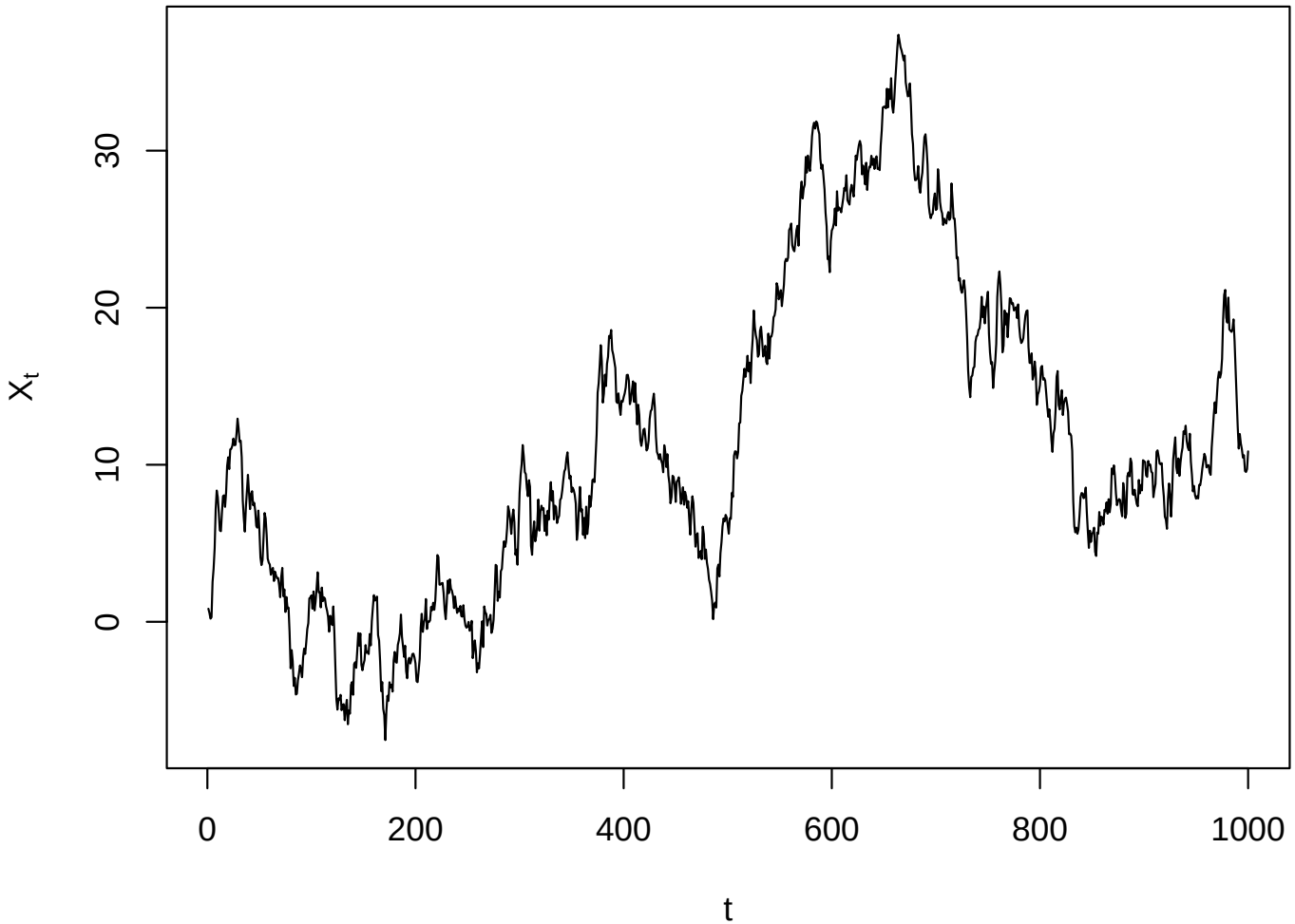
```
set.seed(123456)
n<-1000
z1<-rnorm(n)
z2<-rnorm(n)
x<-cumsum(z1);
y<-0.6*x+z2
lr.reg<-lm(y~x)
error<-residuals(lr.reg)
error.lagged<-error[-c(n-1,n)]
xstar<-diff(x);ystar<-diff(y)
diff.dat<-data.frame(embed(cbind(xstar,ystar),2))
colnames(diff.dat)<-c("xstar","ystar","xstar.1","ystar.1")
```

Simulated Example: R code

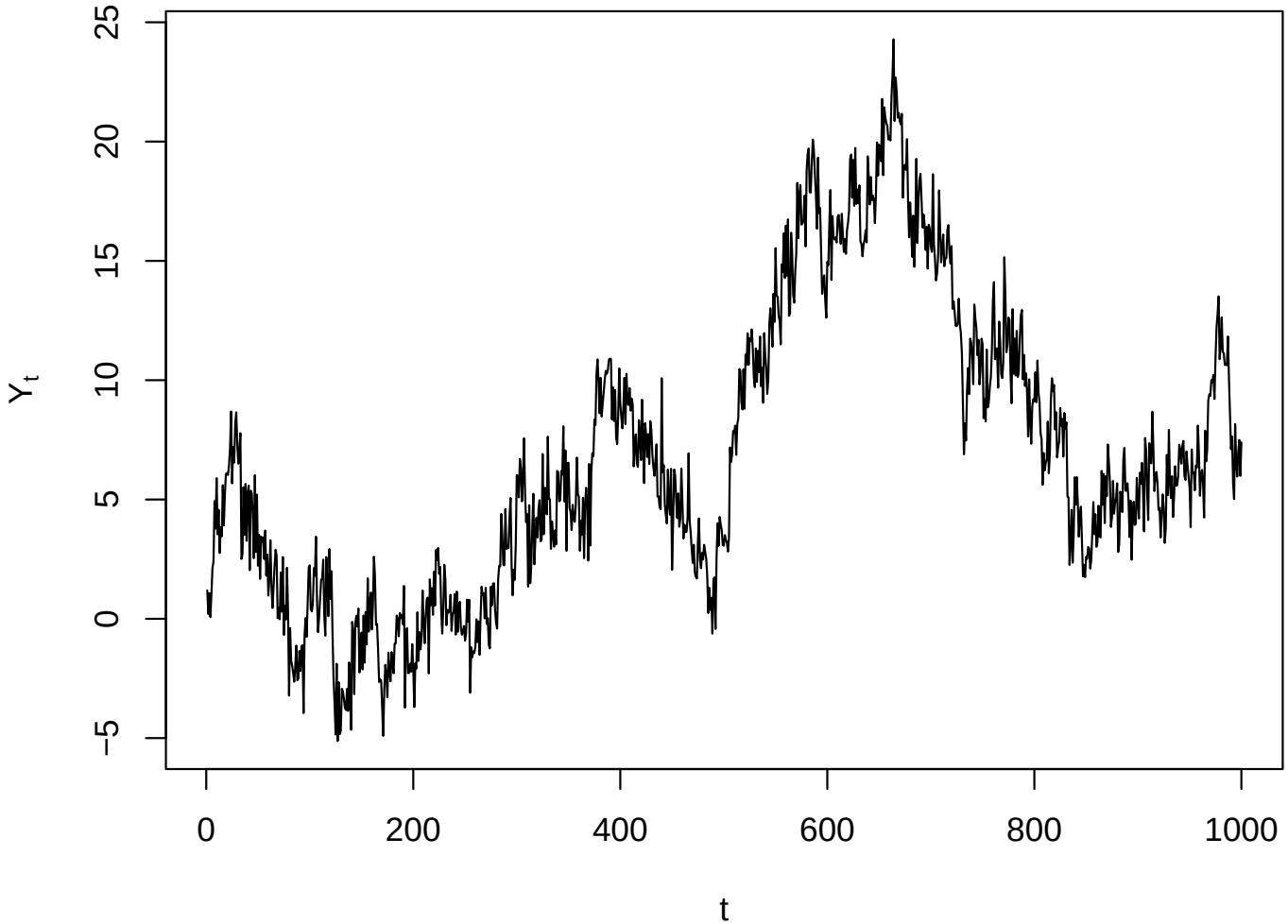
```
ecm.reg<-lm(ystar~error.lagged+xstar.1+ystar.1,data=diff.dat)
summary(ecm.reg)

##
## Call:
## lm(formula = ystar ~ error.lagged + xstar.1 + ystar.1, data = diff.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1545 -0.8358 -0.0354  0.8193  3.5954
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.005528   0.037437   0.148    0.883
## error.lagged -0.936862   0.051442 -18.212   <2e-16 ***
## xstar.1       0.602592   0.043770  13.767   <2e-16 ***
## ystar.1      -0.972102   0.036773 -26.436   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.183 on 994 degrees of freedom
## Multiple R-squared:  0.4129, Adjusted R-squared:  0.4112
## F-statistic: 233.1 on 3 and 994 DF,  p-value: < 2.2e-16
```

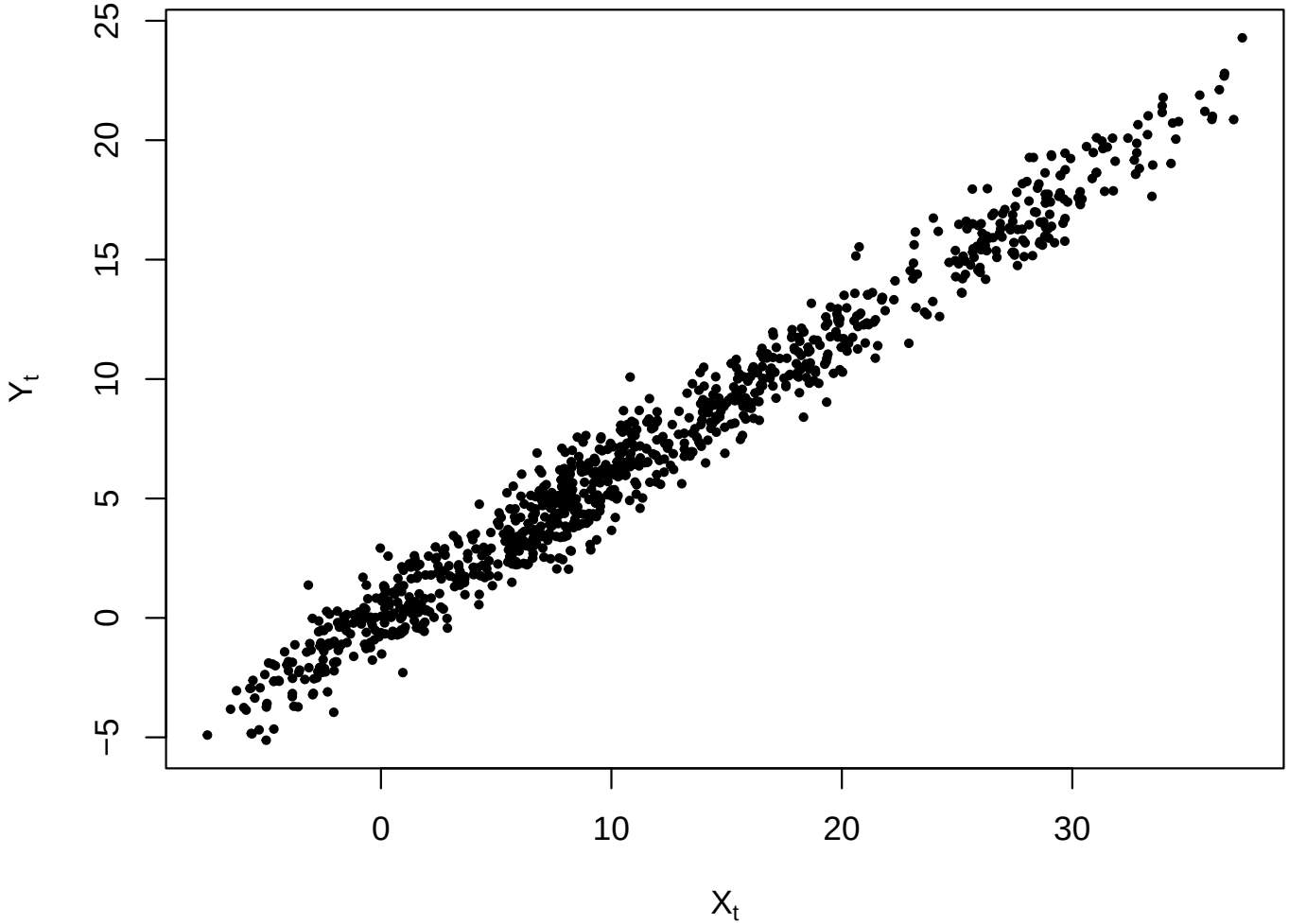
Simulated Example: R code



Simulated Example: R code



Simulated Example: R code



Results

Example: Cointegrating series

```
##                Estimate Std. Error    t value Pr(>|t|)
## (Intercept)    0.005528   0.037437   0.147668 0.882635
## error.lagged -0.936862   0.051442 -18.212078 0.000000
## xstar.1        0.602592   0.043770  13.767172 0.000000
## ystar.1       -0.972102   0.036773 -26.435535 0.000000
```

That is

$$\widehat{\psi}_0 = 0.0055 \quad \widehat{\gamma}_1 = -0.9369 \quad \widehat{\psi}_{11} = 0.6026 \quad \widehat{\psi}_{21} = -0.9721$$

5.5 State Space Models

A state space model can better represent the actual dynamics of a data generation process.

The principle is that an observed series is in fact generated as the marginal process derived from a multivariate time series model, structured so that the model is specified via two equations

- the state equation
- the observation equation

5.5 State Space Models

The general (Markov) state equation takes the form

$$\mathbf{X}_t = \mathbf{g}(\mathbf{x}_{t-1}; \theta_{t-1}) + \mathbf{W}_t$$

and the general observation equation takes the form

$$\mathbf{Y}_t = \mathbf{f}(\mathbf{x}_t; \theta_t) + \mathbf{V}_t.$$

The process $\{\mathbf{Y}_t\}$ is observed, the process $\{\mathbf{X}_t\}$ is latent or unobserved, and one focus of inferential interest.

Let $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$.

Inference in General State Space Models

Inference goals for state space models are

- filtering/tracking: estimate state at time t from observations up to time t

$$p(\mathbf{x}_t | \mathbf{y}_{1:t})$$

- smoothing: estimate state at time t from past and possibly some future observations

$$p(\mathbf{x}_t | \mathbf{y}_{1:(t+s)})$$

- prediction/forecasting: estimate state at time t from observations up to a previous time point $t - s$

$$p(\mathbf{x}_t | \mathbf{y}_{1:(t-s)})$$

Inference in General State Space Models

The inference mechanism is probabilistic; we compute the quantities of interest sequentially in the following recursive calculation:

- Prediction

$$p(\mathbf{x}_t|\mathbf{y}_{1:(t-1)}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:(t-1)})\ d\mathbf{x}_{t-1}$$

- Filtering

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:(t-1)})}{p(\mathbf{y}_t|\mathbf{y}_{1:(t-1)})}$$

Initially, we assume that the system parameters θ_t are known.

Inference in General State Space Models

The above equations have an analytical solution if all densities in the state and observation equation are Gaussian, and both the state and the observation equation are linear.

- If these conditions are met, the Kalman filter provides the optimal (Bayesian) solution to the tracking problem.
- If these conditions are not met:
  - Extended Kalman filter: analytical approximation.
  - Particle filter: sampling-based approximation.

Linear Gaussian State Space Models

The general form of a linear Gaussian state space model is given by the equations

$$\mathbf{Y}_t = \mathbf{F}_t\mathbf{X}_t + \mathbf{d}_t + \mathbf{V}_t \tag{2a}$$

$$\mathbf{X}_{t+1} = \mathbf{G}_t\mathbf{X}_t + \mathbf{e}_t + \mathbf{W}_t \tag{2b}$$

where

- $\mathbf{Y}_t$ is $(K \times 1)$
- $\mathbf{X}_t$ is $(L \times 1)$.

Linear Gaussian State Space Models

In (2)

- $\mathbf{F}_t$ is $(K \times L)$,
- $\mathbf{d}_t$ is $(K \times 1)$,
- $\mathbf{G}_t$ is $(L \times L)$,
- $\mathbf{e}_t$ is $(L \times 1)$.

These system matrices are assumed to be non-stochastic.

Linear Gaussian State Space Models

We may write

$$\begin{bmatrix} \mathbf{V}_t \\ \mathbf{W}_t \end{bmatrix} = \mathbf{H}_t \mathbf{Z}_t$$

where

$$\mathbf{Z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

is a Gaussian IID process, and $\mathbf{H}_t$ induces covariance between the components of $\mathbf{V}_t$ and $\mathbf{W}_t$.

Linear Gaussian State Space Models

The formulation is completed by the assumption that the initial state vector $\mathbf{X}_1$ is independent of $\mathbf{Z}_t$ at all times t and has unconditional mean and variance $\mathbf{a}_{1|0}$ and $\mathbf{P}_{1|0}$ respectively.

The state space model is characterized by the properties that the system matrices, the disturbance terms $\{\mathbf{Z}_t\}$ and initial state vector $\mathbf{X}_1$ possess.

If the system matrices do not evolve with time, the state space model is called time-invariant or time-homogeneous.

Linear Gaussian State Space Models

Example: AR(2) Model

Let $\{X_t\}$ be a zero-mean, Gaussian AR(2) process given by

$$X_{t+1} = \phi_1 X_t + \phi_2 X_{t-1} + W_t$$

where $W_t \sim \mathcal{N}\left(0, \sigma_W^2\right)$.

Suppose also that we observe the process $\{Y_t\}$ with noise, so that

$$Y_t = X_t + V_t$$

with $V_t \sim \mathcal{N}\left(0, \sigma_V^2\right)$ and independent of W_t.

Linear Gaussian State Space Models

Example: AR(2) Model

A state space representation has $K = 1$ and $L = 2$

$$Y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} \sigma_V & 0 \end{bmatrix} \begin{bmatrix} Z_{1,t} \\ Z_{2,t} \end{bmatrix}$$

$$\begin{bmatrix} X_{t+1} \\ X_t \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} 0 & \sigma_W \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z_{1,t} \\ Z_{2,t} \end{bmatrix}$$

where $\mathbf{Z}_t = (Z_{1,t}, Z_{2,t})^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$.

The state space model is time-homogeneous and Gaussian, and the observation and state equations are uncorrelated.

5.6 Kalman Filter

In the linear Gaussian case, the Kalman filter is used for

- Filtering: computes for all t

$$\mathbb{E}[\mathbf{X}_t|\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}] \qquad \text{Var}[\mathbf{X}_t|\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}].$$

- Smoothing: computes for all $t < n$

$$\mathbb{E}[\mathbf{X}_t|\mathbf{Y}_{1:n} = \mathbf{y}_{1:n}] \qquad \text{Var}[\mathbf{X}_t|\mathbf{Y}_{1:n} = \mathbf{y}_{1:n}].$$

- Prediction: computes for all t

$$\mathbb{E}[\mathbf{X}_{t+1}|\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}] \qquad \text{Var}[\mathbf{X}_{t+1}|\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}].$$

The Kalman filter consists of a set of recursions that compute these quantities.

5.6 Kalman Filter

For simplicity, we assume that

$$\mathbf{d}_t = \mathbf{e}_t = \mathbf{0}$$

and assume that the processes $\{\mathbf{V}_t\}$ and $\{\mathbf{W}_t\}$ are uncorrelated, with

$$\text{Var}[\mathbf{V}_t] = \mathbf{\Sigma}_t \qquad \text{Var}[\mathbf{W}_t] = \mathbf{\Omega}_t.$$

5.6 Kalman Filter

The Kalman filter recursively computes the quantities

- $\mathbf{a}_{t|t} = \mathbb{E}[\mathbf{X}_t|\mathbf{y}_{1:t}]$,
- $\mathbf{a}_{t+1|t} = \mathbb{E}[\mathbf{X}_{t+1}|\mathbf{y}_{1:t}]$,
- $\mathbf{P}_{t|t} = \mathrm{MSE}[\mathbf{X}_t|\mathbf{y}_{1:t}]$,
- $\mathbf{P}_{t+1|t} = \mathrm{MSE}[\mathbf{X}_{t+1}|\mathbf{y}_{1:t}]$.

where MSE is the mean square error which is equivalent to the one-step ahead prediction variance.

5.6 Kalman Filter

Then, starting with $\mathbf{a}_{1|0}$ and $\mathbf{P}_{1|0}$, $\mathbf{a}_{t|t}$, $\mathbf{a}_{t+1|t}$ and their MSEs are obtained by running for $t = 1, \ldots, n$, the recursions are

$$\mathbf{r}_t = \mathbf{y}_t - \mathbf{F}_t \mathbf{a}_{t|t-1} \tag{3a}$$

$$\mathbf{M}_t = \mathbf{F}_t \mathbf{P}_{t|t-1} \mathbf{F}_t^\top + \mathbf{\Sigma}_t \tag{3b}$$

$$\mathbf{a}_{t|t} = \mathbf{a}_{t|t-1} + \mathbf{P}_{t|t-1} \mathbf{F}_t^\top \mathbf{M}_t^{-1} \mathbf{r}_t \tag{3c}$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1} \mathbf{F}_t^\top \mathbf{M}_t^{-1} \mathbf{F}_t \mathbf{P}_{t|t-1} \tag{3d}$$

$$\mathbf{a}_{t+1|t} = \mathbf{G}_t \mathbf{a}_{t|t}, \tag{3e}$$

$$\mathbf{P}_{t+1|t} = \mathbf{G}_t \mathbf{P}_{t|t} \mathbf{G}_t^\top + \mathbf{\Omega}_t \tag{3f}$$

5.6 Kalman Filter

Notes:

- quantities $\mathbf{r}_t$ and $\mathbf{M}_t$ in (3b) are the one-step-ahead error in forecasting $\mathbf{Y}_t$ conditional on the information set at time $t-1$ and its MSE. This facilitates computation of the likelihood function.
- the quantities $\mathbf{a}_{t|t}$ and $\mathbf{a}_{t|t-1}$ are optimal estimators of $\mathbf{X}_t$ conditional on the available information, in terms of minimum mean square.
- the latter property holds under Gaussianity, otherwise, $\mathbf{a}_{t|t}$ and $\mathbf{a}_{t|t-1}$ are optimal only within the class of linear estimators (the minimum mean square linear estimators) of $\mathbf{X}_t$ conditional on $\mathbf{y}_{1:t}$ and $\mathbf{y}_{1:(t-1)}$.

5.6 Kalman Filter

- Equations (3c) and (3e) can be combined to yield recursions that compute the one-step-ahead prediction estimates of $\mathbf{X}_{t+1}$ given $\mathbf{y}_{1:t}$.

  Similarly, by taking together (3d) and (3f), a single set of recursions for the MSE is obtained, which goes directly from $\mathbf{P}_{t|t-1}$ to $\mathbf{P}_{t+1|t}$.

- The resulting recursions, along with (3b), are

$$\begin{aligned}
\mathbf{a}_{t+1|t} &= \mathbf{G}_t \mathbf{a}_{t|t-1} + \mathbf{K}_t \mathbf{r}_t && \text{(4a)} \\
\mathbf{K}_t &= \mathbf{G}_t \mathbf{P}_{t|t-1} \mathbf{F}_t^\top \mathbf{M}_t^{-1} && \text{(4b)} \\
\mathbf{P}_{t+1|t} &= \mathbf{G}_t \mathbf{P}_{t|t-1} \mathbf{L}_t^\top + \mathbf{\Omega}_t && \text{(4c)} \\
\mathbf{L}_t &= \mathbf{F}_t - \mathbf{K}_t \mathbf{G}_t. && \text{(4d)}
\end{aligned}$$

Parameter Estimation

The Kalman filter may be used for estimation of unknown parameters $\boldsymbol{\theta}$ that appear in the system matrices.

The likelihood for data $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ can be constructed as

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_T) = p\left(\mathbf{y}_n | \mathbf{y}_{(n-1)}\right) \cdots p(\mathbf{y}_2 | \mathbf{y}_1) \, p(\mathbf{y}_1)$$

$$= \prod_{t=1}^{n} p(\mathbf{y}_t | \mathbf{y}_{t-1})$$

Parameter Estimation

For a Gaussian model, by taking conditional expectations on both sides of the observation equation, with $\mathbf{d}_t \equiv 0$, then

$$\mathbb{E}\left[\mathbf{Y}_t \middle| \mathbf{y}_{1:(t-1)}\right] = \mathbf{F}_t \mathbf{a}_{t|t-1} \qquad \text{Var}\left[\mathbf{Y}_t \middle| \mathbf{y}_{1:(t-1)}\right] = \mathbf{M}_t.$$

The one-step-ahead prediction density is a multivariate normal distribution

$$p\left(\mathbf{y}_t \middle| \mathbf{y}_{1:(t-1)}\right) \equiv \mathcal{N}\left(\mathbf{F}_t \mathbf{a}_{t|t-1}, \mathbf{M}_t\right).$$

Parameter Estimation

Thus, the log-likelihood function is given by

$$-\frac{1}{2}\sum_{t=1}^{n}\log\left(\det\mathbf{M}_t\right)-\frac{1}{2}\sum_{t=1}^{n}\mathbf{r}_t^{\top}\mathbf{M}_t^{-1}\mathbf{r}_t$$

where as before

$$\mathbf{r}_t=\left(\mathbf{y}_t-\mathbf{F}_t\mathbf{a}_{t|t-1}\right)$$

Typically, numerical procedures are used in order to maximize the log-likelihood to obtain the ML estimates of the parameters $\boldsymbol{\theta}$ which are consistent and asymptotically normal.

Smoothing

We may compute

$$\mathbf{a}_{t|n} = \mathbb{E}\big[\mathbf{X}_t \big| \mathbf{y}_{1:n}\big]$$

along with its MSE, $\mathbf{P}_{t|n}$ recursively, for all $t = 1, \ldots, n-1$.

The latter quantities are computed via a set of backward recursions and use the quantities $\mathbf{a}_{t+1|t}$, $\mathbf{a}_{t|t}$ and the MSE matrices $\mathbf{P}_{t|t}$, $\mathbf{P}_{t+1|t}$, which are obtained from (3).

Smoothing

In particular, to obtain $\mathbf{a}_{t|n}$ and $\mathbf{P}_{t|n}$, we start with $\mathbf{a}_{n|n}$ and $\mathbf{P}_{n|n}$ and we run backwards for $t = n-1, \ldots, 0$

$$\mathbf{a}_{t|n} = \mathbf{a}_{t|t} + \mathbf{P}_t^* \left(\mathbf{a}_{t+1|n} - \mathbf{a}_{t+1|t} \right)$$

$$\mathbf{P}_{t|n} = \mathbf{P}_{t|t} + \mathbf{P}_t^* \left(\mathbf{P}_{t+1|n} - \mathbf{P}_{t+1|t} \right) \mathbf{P}_t^*$$

$$\mathbf{P}_t^* = \mathbf{P}_{t|t} \mathbf{G}_t^\top \mathbf{P}_{t+1|t}^{-1}.$$

Non-Linear Non-Gaussian Models

If the state space model is not Gaussian, the likelihood can still be constructed in the same way using the minimum mean square linear estimators of the state vector.

However, the estimators that maximize the likelihood are the quasi-maximum likelihood (QML) estimators of the parameters, which are also consistent and asymptotically normal.

Particle Filtering

The particle filter uses Monte Carlo methods, in particular importance sampling, to construct the approximations. If

$$\mathbb{E}_f[g(X)] = \int g(x)f(x)\ dx$$

then

$$\mathbb{E}_f[g(X)] \equiv \mathbb{E}_{f_0}\left[\frac{f(X)g(X)}{f_0(X)}\right] = \int \frac{g(x)f(x)}{f_0(x)} f_0(x)\ dx$$

for some importance sampling density f_0.

Particle Filtering

We then approximate the integral by sampling $x_1, \ldots, x_N$ from f_0 for large N, and constructing the approximation

$$\widehat{\mathbb{E}}_{f_N} = \frac{1}{N} \sum_{i=1}^{N} \frac{g(x_i) f(x_i)}{f_0(x_i)} = \frac{1}{N} \sum_{i=1}^{N} w_i g(x_i),$$

say, where, if f_0 is chosen carefully

$$\widehat{\mathbb{E}}_{f_N} \longrightarrow \mathbb{E}_f[g(X)]$$

as $N \longrightarrow \infty$.

Particle Filtering

If the function f is known only up to proportionality, the same procedure can work if the weights w_i are suitably normalized to sum to one, that is, if above

$$w_i = \frac{f(x_i)}{f_0(x_i)}$$

is replaced by

$$w_i^\star = \frac{w_i}{W} \qquad \text{with} \qquad W = \sum_{i=1}^{n} w_i .$$

as W approximates the integral

$$\int f(x)\ dx = \int \frac{f(x)}{f_0(x)} f_0(x)\ dx$$

Particle Filtering

Given a non-linear and non-Gaussian state space model, the goal is to compute the filtering and prediction densities

$$p\left(\mathbf{x}_t \middle| \mathbf{y}_{1:t}\right) \qquad p\left(\mathbf{x}_{t+1} \middle| \mathbf{y}_{1:t}\right)$$

at each time t via a Monte Carlo approximation.

We aim to maintain a sample of particles

$$\left(x_{t1}, \ldots, x_{tN}\right)$$

for each t that can be used to approximate these distributions.

Particle Filtering

At each time point t, we perform the

Sequential Importance Sampling

(or SIS) procedure:

Particle Filtering

- Draw random samples from a chosen importance density that can be sampled directly.
- Compute a normalized weight for each particle
- Approximate the target density at time t by the weighted sum of the particles.

Given sufficiently large numbers of particles, this IS characterization converges to the target density.

Particle Filtering

In the filtering problem, we approximate $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ by

- particles $\mathbf{x}_t = (x_{t1}, \ldots, x_{tN})$, sampled from density f_0.
- normalized weights $\boldsymbol{w}_t^\star = (w_{t1}^\star, \ldots, w_{tN}^\star)$

via the discrete distribution $\{(w_{ti}^\star, x_{ti}), i = 1 \ldots, N\}$.

Particle Filtering

Two key questions are

- What importance density f_0 can or should be chosen?
- How do we compute the weights efficiently (in a recursive fashion) such that for a given weight $w_{ti}^{\star}$, we just need a new observation $\mathbf{y}_t$ to get an approximation to $p(\mathbf{x}_t|\mathbf{y}_{1:t})$.

A simple but effective choice for f_0 sampling density at time t is the transition density: $f_0(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{1:t}) \equiv p(\mathbf{x}_t|\mathbf{x}_{t-1})$ as then

$$w_{ti}^{\star} \propto w_{t-1,i}^{\star} p(\mathbf{y}_t|\mathbf{x}_{ti}^{\star}).$$

Particle Degeneracy

A problem with sequential importance sampling is that typically, after a few iterations, all but one particle have negligible weights. This phenomenon is termed degeneracy.

The degree of degeneracy is represented by the quantity

$$N_t = \left(\sum_{i=1}^{N} (w_{ti}^{\star})^2 \right)^{-1}$$

which is the effective sample size.

Particle Degeneracy

Possible solutions to overcome degeneracy effects:

- brute force: very large N
- optimized choice of importance density
- resampling

Resampling

The goal of resampling is to eliminate particles with small weights and replace by new particles drawn from the vicinity of particles with larger weights

- concentrate on particles with large weights
- model important parts of the distribution more precisely

Various efficient algorithms exist for resampling.

- eg Sampling Importance Resampling (SIR)

5.7 Hidden Markov Models

If the unobserved state X_t is **discrete**, then state space models are often called hidden Markov models (HMMs).

In this case, the state equation is replaced by a Markov chain model: if $\{X_t\}$ can take values on $\{1, 2, \ldots, S\}$ say, then for $t > 1$

$$p(x_t | x_{1:(t-1)}) \equiv p(x_t | x_{t-1})$$

where

$$\Pr[X_t = l | X_{t-1} = k] = p_{kl}$$

The $S \times S$ transition matrix $\mathbf{P}$, along with the (marginal) distribution of X_1 ($\Pr[X_1 = k], k = 1, \ldots, S$) characterize the nature of the state evolution. Note that $\mathbf{P}$ is a stochastic matrix whose rows sum to 1.

5.7 Hidden Markov Models

All of the previous types of calculation

- filtering $p(x_t|y_{1:t})$
- prediction $p(x_t|y_{1:(t-1)})$
- smoothing $p(x_t|y_{1:n})$
- parameter estimation

are required. The main difference is that the filtering, prediction and smoothing distributions are **discrete**, and therefore somewhat more straightforward to compute.

The observation variable y_t can be either discrete or continuous.

5.7 Hidden Markov Models

Example: $S = 2$

Suppose for $t = 1, 2, \ldots, n$

$$Y_t = \mu(X_t) + \epsilon_t$$

and $\{X_t\}$ follows a Markov chain with transition matrix

$$\mathbf{P} = \begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix}$$

and the distribution of X_1 is the stationary (or equilibrium) distribution, $\boldsymbol{\pi}$, of the chain defined as the solution of

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$$

Example: $S = 2$

We compute $\boldsymbol{\pi}$ (a 1×2 vector) by

$$[1,1]^{\mathsf{T}} \boldsymbol{\pi} = \lim_{t \longrightarrow \infty} \mathbf{P}^t = \begin{bmatrix} 4/11 & 7/11 \\ 4/11 & 7/11 \end{bmatrix}$$

so $\boldsymbol{\pi} = (4/11, 7/11)$.

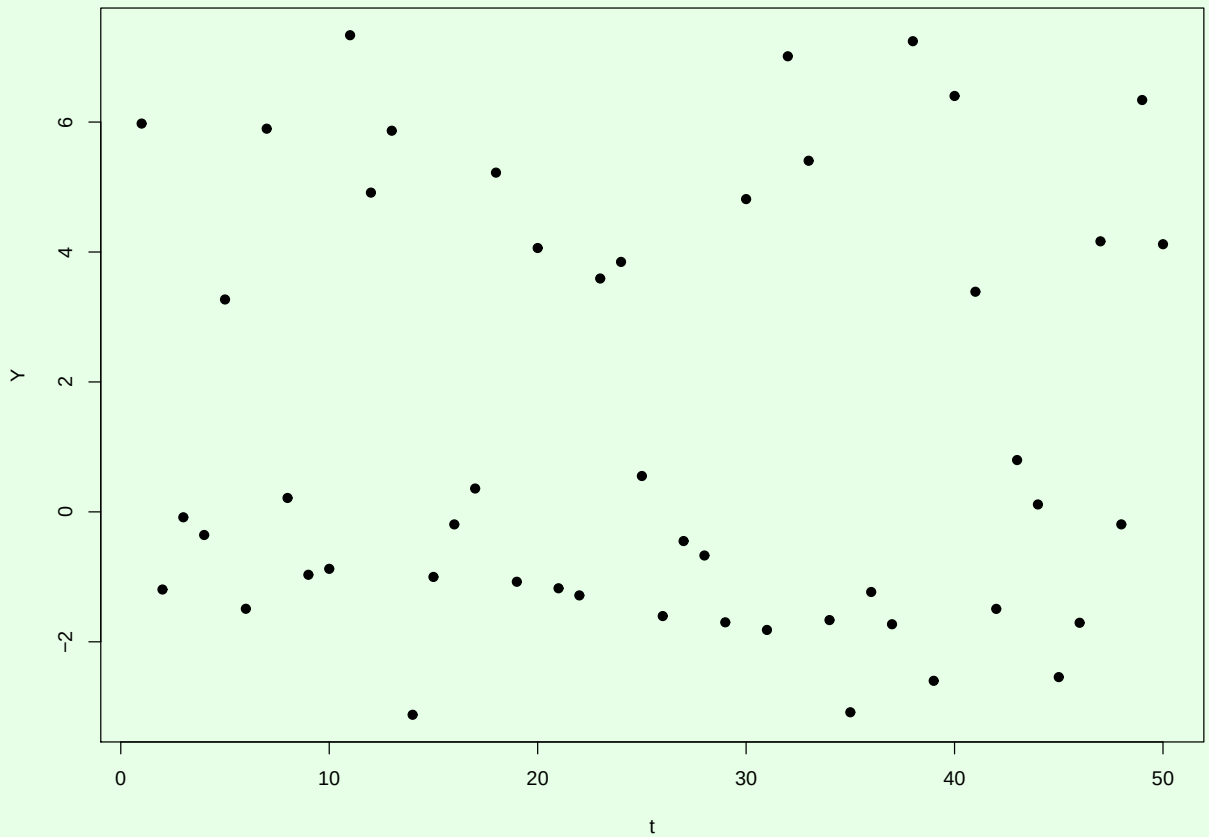
Suppose

$$\mu(x) = \begin{cases} \mu_1 & x = 1 \\ \mu_2 & x = 2 \end{cases}$$

and $\{\epsilon_t\} \sim \mathcal{N}(0,1)$, IID.

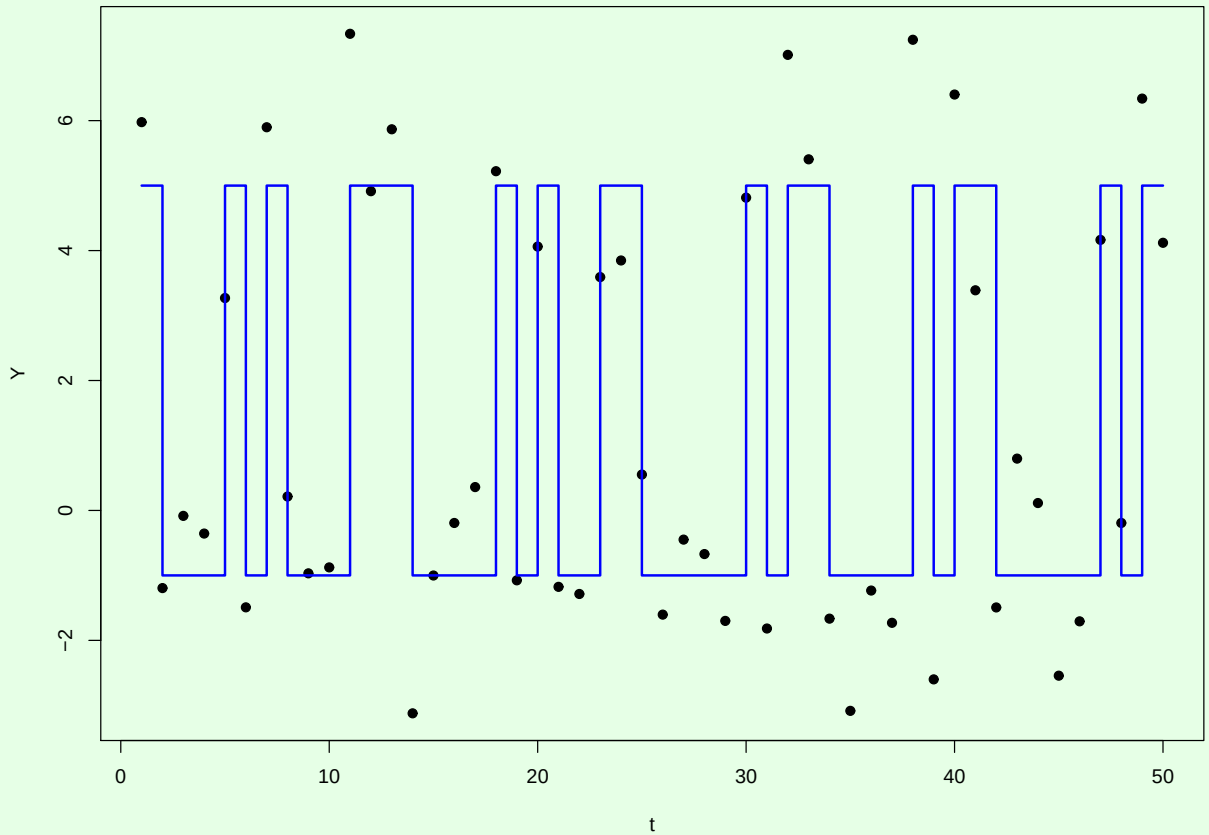
5.7 Hidden Markov Models

Example: $S = 2$

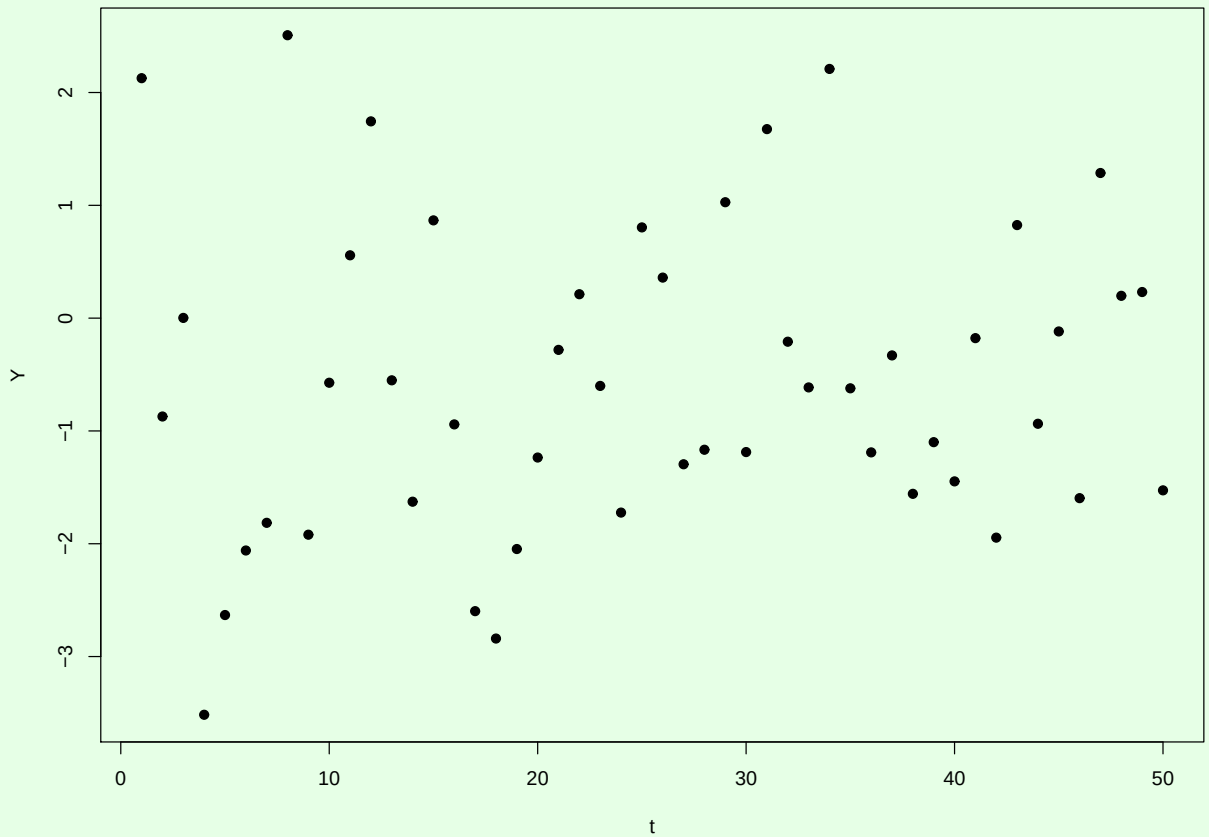


5.7 Hidden Markov Models

Example: $S = 2, \mu_1 = 5, \mu_2 = -1$

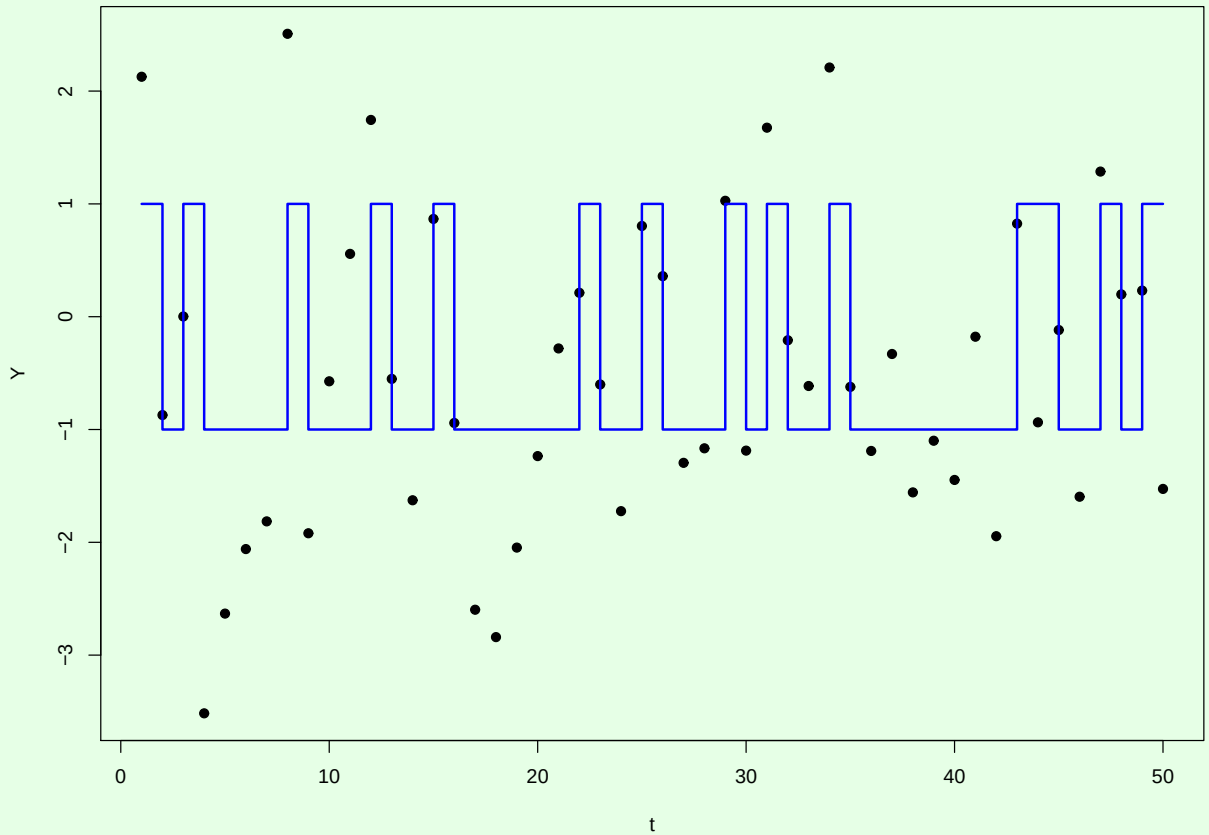


Example: $S = 2$



5.7 Hidden Markov Models

Example: $S = 2, \mu_1 = 1, \mu_2 = -1$



5.7 Hidden Markov Models

Filtering: By Bayes theorem

$$
\begin{aligned}
p(x_t|y_{1:t}) &= p(x_t|y_t, y_{1:(t-1)}) \\
&= \frac{p(y_t|x_t, y_{1:(t-1)})p(x_t|y_{1:(t-1)})}{p(y_t|y_{1:(t-1)})} \\
&= \frac{p(y_t|x_t)p(x_t|y_{1:(t-1)})}{\sum_{x=1}^{S} p(y_t|x)p(x|y_{1:(t-1)})} \\
&\propto p(y_t|x_t)p(x_t|y_{1:(t-1)})
\end{aligned}
$$

This is a discrete distribution on $\{1, \ldots, S\}$ that we can compute exactly.

5.7 Hidden Markov Models

Prediction: By the theorem of total probability

$$p(x_{t+1}|y_{1:t}) = \sum_{x=1}^{S} p(x_{t+1}|x)p(x|y_{1:t})$$

- $p(x_{t+1}|x)$ is determined the xth row of $\mathbf{P}$, for each $x = 1, 2, \ldots, S$
- $p(x|y_{1:t})$ is the previous filtering distribution.

5.7 Hidden Markov Models

Smoothing: By the theorem of total probability, we can compute backwards $t = n, n-1, \ldots, 1$,

$$p(x_t|y_{1:n}) = \sum_{x_{t+1}=1}^{S} p(x_t|x_{t+1}, y_{1:n}) p(x_{t+1}|y_{1:n})$$

$p(x_{t+1}|y_{1:n})$ is the previous smoothing distribution.

For the first term

$$
\begin{aligned}
p(x_t|x_{t+1}, y_{1:n}) &\equiv p(x_t|x_{t+1}, y_{1:t}) \\
&= \frac{p(x_t|y_{1:t})p(x_{t+1}|x_t, y_{1:t})}{p(x_{t+1}|y_{1:t})} \\
&= \frac{p(x_t|y_{1:t})p(x_{t+1}|x_t)}{p(x_{t+1}|y_{1:t})}
\end{aligned}
$$

which can be computed from the filtering and prediction distributions.

Discrete HMMs

The HMM structure is commonly implemented when the observation is also discrete.

As usual let $x_{1:n}$ denote the hidden state sequence, and $y_{1:n}$ denote the observed data sequence.

Discrete HMMs

Then the model is specified by

- transition probabilities

$$p_{kl} = \Pr\left[X_t = l | X_{t-1} = k\right]$$

  with initial distribution $\Pr[X_1 = k]$, $k = 1, \ldots, S$.

- emission probabilities

$$e_k(m) = \Pr\left[Y_t = m | X_t = k\right]$$

$e_k(m)$ can be specified for a finite set ($m \in \{1, 2, \ldots, M\}$), or on the integers.

Discrete HMMs: Inference Tasks

We seek the most probable state path, that is, the most probable sequence of states in light of the data:

$$\begin{aligned} x^*_{1:n} &= \arg\max_{x_{1:n}} p\left(x_{1:n}, y_{1:n}\right) \\ &= \arg\max_{x_{1:n}} p\left(x_{1:n} \middle| y_{1:n}\right) \end{aligned}$$

- $p\left(x_{1:n}, y_{1:n}\right)$ is the joint probability or joint likelihood
- $p\left(x_{1:n} \middle| y_{1:n}\right)$ is the posterior probability

Discrete HMMs: The Viterbi Algorithm

(i) Initialize vector $v_k(0) = \Pr[X_1 = k]$, $k = 1, \ldots, S$.

(ii) Recursion: for $t = 1, ..., n$, compute

$$v_l(t) = e_l(y_t) \max_k (v_k(t-1) p_{kl})$$

$$point_t(l) = \arg\max_k (v_k(t-1) p_{kl})$$

for each $l = 1, 2, \ldots, S$

Discrete HMMs: The Viterbi Algorithm

(iii) Termination

$$p\left(y_{1:n}, x^*_{1:n}\right) = \max_k \left(v_k(n)\right)$$

$$x^*_n = \arg\max_k \left(v_k(n)\right)$$

(iv) Trace back the optimal path

$$x^*_{t-1} = point_t\left(x^*_t\right) \qquad t = n, n-1, \ldots, 2.$$

Discrete HMMs: The Forward Algorithm

The probability of an observed sequence (or likelihood) is

$$p\left(y_{1:n}\right) = \sum_{x_{1:n}} p\left(y_{1:n}, x_{1:n}\right)$$

However, the direct computation of this sum is not possible, so a recursive calculation is needed.

Discrete HMMs: The Forward Algorithm

(i) Initialize vector $f_k(1) = e_k(y_1)\pi_k$ for $k = 1, 2 \ldots, S$

(ii) Recursion: for $t = 2, ..., n,$ and $l = 1, 2, \ldots, S,$

$$f_l\left(t\right) = e_l\left(y_t\right) \sum_{k=1}^{S} f_k\left(t-1\right) p_{kl}$$

(iii) Final calculation:

$$p\left(y_{1:n}\right) = \sum_{k=1}^{S} f_k\left(n\right).$$

Discrete HMMs: The Backward Algorithm

We seek the posterior (smoothing) probabilities

$$p\left(X_t = k | y_{1:n}\right) = \frac{p\left(X_t = k, y_{1:n}\right)}{p\left(y_{1:n}\right)}$$

$$= \frac{p\left(y_1, ..., y_t, X_t = k\right) p\left(y_{t+1}, ..., y_n | X_t = k\right)}{p\left(y_{1:n}\right)}$$

for $k = 1, 2, \ldots, S$.

Direct computation is not possible, so a recursive calculation is needed, after completing the Forward Algorithm.

Discrete HMMs: The Backward Algorithm

The backward algorithm works in reverse order through the data:

(i) Initialize element $b_k(n) = f_k(n)$ for $k = 1, \ldots, S$.

(ii) Backwards recursion: for $t = n-1, ..., 1$ compute

$$b_k(t) = \sum_{l=1}^{S} p_{kl}\, e_l(y_{t+1})\, b_l(t+1)$$

for $k = 1, 2, \ldots, S$.

Discrete HMMs: The Backward Algorithm

The forward and backward probabilities, $f_k(t)$ and $b_k(t)$, are in fact the probabilities

$$f_k(t) = p\left(y_1, ..., y_t, X_t = k\right)$$
$$b_k(t) = p\left(y_{t+1}, ..., y_n | X_t = k\right)$$

Thus we have

$$p\left(X_t = k, y_{1:n}\right) = f_k(t) b_k(t)$$

and hence, for $k = 1, \ldots, S$,

$$p\left(X_t = k | y_{1:n}\right) = \frac{f_k(t) b_k(t)}{p\left(y_{1:n}\right)}$$

Discrete HMMs: Estimation

The algorithms above are all expressed in terms of known transition and emission probabilities. These probabilities can be pre-specified, estimated from observed sequences, or estimated during the state sequence reconstruction.

If the state sequence is known, let n_{kl} be the number of transitions from state k to state l, and let $r_k(b)$ be the number of times character b is emitted from state k. Then we may estimate the transition and emission probabilities as follows

$$\hat{p}_{kl} = \frac{n_{kl}}{\sum_{l'} n_{kl'}} \qquad \hat{e}_k(b) = \frac{r_k(b)}{\sum_{b'} r_k(b')} \tag{5}$$

The Baum-Welch Algorithm

The Baum-Welch algorithm estimates n_{kl} and $r_k(b)$ as the expected number of transitions and emissions, which are computed using the forward and backward algorithms.

The Baum-Welch Algorithm

The probability of a transition $k \to l$ at time t is

$$p\left(X_t = k, X_{t+1} = l \middle| y_{1:n}\right) = \frac{f_k\left(t\right) p_{kl} e_l\left(y_{t+1}\right) b_l\left(t+1\right)}{p\left(y_{1:n}\right)}$$

Suppose that N replicate observed sequences

$$y_{1:n}^1, \ldots, y_{1:n}^N$$

are available. Then the expected number of times this transition will be observed (in light of the data) is

$$n_{kl} = \sum_{i=1}^{N} \left\{ \frac{\sum_t f_k^i\left(t\right) a_{kl} e_l(y_{t+1}^i) b_l^i\left(t+1\right)}{p(y_{1:n}^i)} \right\}$$

The Baum-Welch Algorithm

In this formula

- $y^i_{1:n}$ is the ith observed sequence
- $f^i_k(t)$ is the forward probability from the forward algorithm implemented for sequence i
- $b^i_l(t)$ is the backward probability from the backward algorithm implemented for sequence i

Similarly,

$$r_k(b) = \sum_{i=1}^{N} \left\{ \frac{\sum_{t \in \mathcal{B}_i} f^i_k(t)\, b^i_l(t)}{p\left(y^i_{1:n}\right)} \right\}$$

where $\mathcal{B}_i$ is the set of indices of locations in replicate sequence i where character b is emitted.

The Baum-Welch Algorithm

(i) Initialize parameter values to some arbitrary values

(ii) Recursion: Set $n_{kl}^{(0)}$, $r_k^{(0)}(b)$ to be some small count value, then for $i = 1, ..., N$: calculate $f_k^i(t)$ using the Forward Algorithm, and $b_k^i(t)$ using the Backward Algorithm, then set

$$n_{kl}^{(i)} = n_{kl}^{(i-1)} + \frac{1}{p\left(y_{1:n}^i\right)} \left\{ \sum_{t=1}^{n} f_k^i(i)\, p_{kl}\, e_l\left(y_{t+1}^i\right) b_l^i(t+1) \right\}$$

$$r_k^{(i)}(b) = r_k^{(i-1)}(b) + \frac{1}{p\left(y_{1:n}^i\right)} \left\{ \sum_{t \in \mathcal{B}_i} f_k^i(t) b_k^i(t) \right\}$$

The Baum-Welch Algorithm

(iii) Estimate p_{kl} and $e_k(b)$ using (5), and compute the likelihood

$$\prod_{i=1}^{N} p(y^i_{1:n})$$

(iv) Terminate the recursion if the computed likelihood value is virtually unchanged compared to the previous iteration.

Computation for continuous response HMMs

Note that most of the above calculations carry through to the continuous emission case, that is if the emission probabilities

$$e_k(m) = \Pr[Y_t = m | x_t = k]$$

are replaced by densities

$$e_k(y) = p(y | X_t = k)$$

The forward and backward quantities become unnormalized joint densities

Computation for continuous response HMMs

To avoid underflow in the computation, the computations can either be computed on the log scale, or with continual rescaling of the calculation. For example, in the Viterbi algorithm, the computation of the $v_k(t)$ quantities will underflow quickly without rescaling, but rescaling by

$$c_t = \sum_{k=1}^{S} v_k(t)$$

will prevent this.

The constants c_t should be stored for future use; if the forward and backward probabilities are computed after rescaling, then this rescaling must be undone when, for example, $p(y_{1:n})$ is computed.

Chapter 6
Spectral Methods

6.1 Spectral Analysis of Time Series

The spectral or frequency domain representation of time series $\{X_t\}$ focusses on stationary processes, and constructs an alternative view to the time domain representation studied previously.

The focus on time domain methods is paramount in statistics because observed data are typically collected sequentially in time.

For stationary processes, the non-negative definite autocovariance function admits a spectral decomposition which facilitates alternative inference approaches.

6.2 Spectral Representation

The ACVF of a (zero-mean) stationary process $\{X_t\}$

$$\gamma_X(h) = \mathbb{E}[X_t X_{t+h}]$$

is a symmetric, non-negative definite function of the discrete argument

$$h = 0, \pm 1, \pm 2, \ldots .$$

We exploit the fact that $\gamma_X(h)$ has a unique representation in terms of frequency contributions.

The Spectral Density Function

Let $\{X_t\}$ be stationary and zero-mean, with ACVF $\gamma_X(.)$, with

$$\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty.$$

Then the spectral density function (SDF) of $\{X_t\}$ is the function $f_X(.)$ defined by

$$f_X(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \gamma_X(h) \qquad \lambda \in \mathbb{R}$$

The Spectral Density Function

Note that $f_X(\lambda)$ is well-defined, as $\left|e^{-i\lambda h}\right| = 1$, so

$$\sum_{h=-\infty}^{\infty} |e^{-i\lambda h}\gamma_X(h)| \leqslant \sum_{h=-\infty}^{\infty} |e^{-i\lambda h}||\gamma_X(h)| = \sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty.$$

and $f_X(\lambda)$ is periodic as

$$e^{-i\lambda h} = \cos(\lambda h) - i\sin(\lambda h)$$

so we can restrict attention to $\lambda \in (-\pi, \pi]$

The Spectral Density Function

Example: AR(1)

Consider the stationary AR(1) model

$$(1 - \phi B)X_t = Z_t$$

so that for $h \geqslant 0$

$$\gamma_X(h) = \frac{\sigma_Z^2}{1 - \phi^2}\phi^h$$

The Spectral Density Function

Example: AR(1)

Then

$$f_X(\lambda) = \frac{\sigma_Z^2}{1-\phi^2}\frac{1}{2\pi}\sum_{h=-\infty}^{\infty} e^{-i\lambda h}\phi^{|h|}$$

$$= \frac{\sigma_Z^2}{2\pi(1-\phi^2)}\left[\frac{1}{1-\phi e^{-i\lambda}} + \sum_{h=-\infty}^{-1} e^{-i\lambda h}\phi^{-h}\right]$$

$$= \frac{\sigma_Z^2}{2\pi(1-\phi^2)}\left[\frac{1}{1-\phi e^{-i\lambda}} + \sum_{h=1}^{\infty} e^{i\lambda h}\phi^{h}\right]$$

$$= \frac{\sigma_Z^2}{2\pi(1-\phi^2)}\left[\frac{1}{1-\phi e^{-i\lambda}} + \frac{1}{1-\phi e^{i\lambda}} - 1\right]$$

The Spectral Density Function

Example: AR(1)

Thus

$$f_X(\lambda) = \frac{\sigma_Z^2}{2\pi(1-\phi^2)} \frac{2 - \phi e^{-i\lambda} - \phi e^{i\lambda} - (1-\phi e^{-i\lambda})(1-\phi e^{i\lambda})}{(1-\phi e^{-i\lambda})(1-\phi e^{i\lambda})}$$

$$= \frac{\sigma_Z^2}{2\pi} \frac{1}{(1-\phi e^{-i\lambda})(1-\phi e^{i\lambda})}$$

$$= \frac{\sigma_Z^2}{2\pi} \frac{1}{|1-\phi e^{-i\lambda}|^2}$$

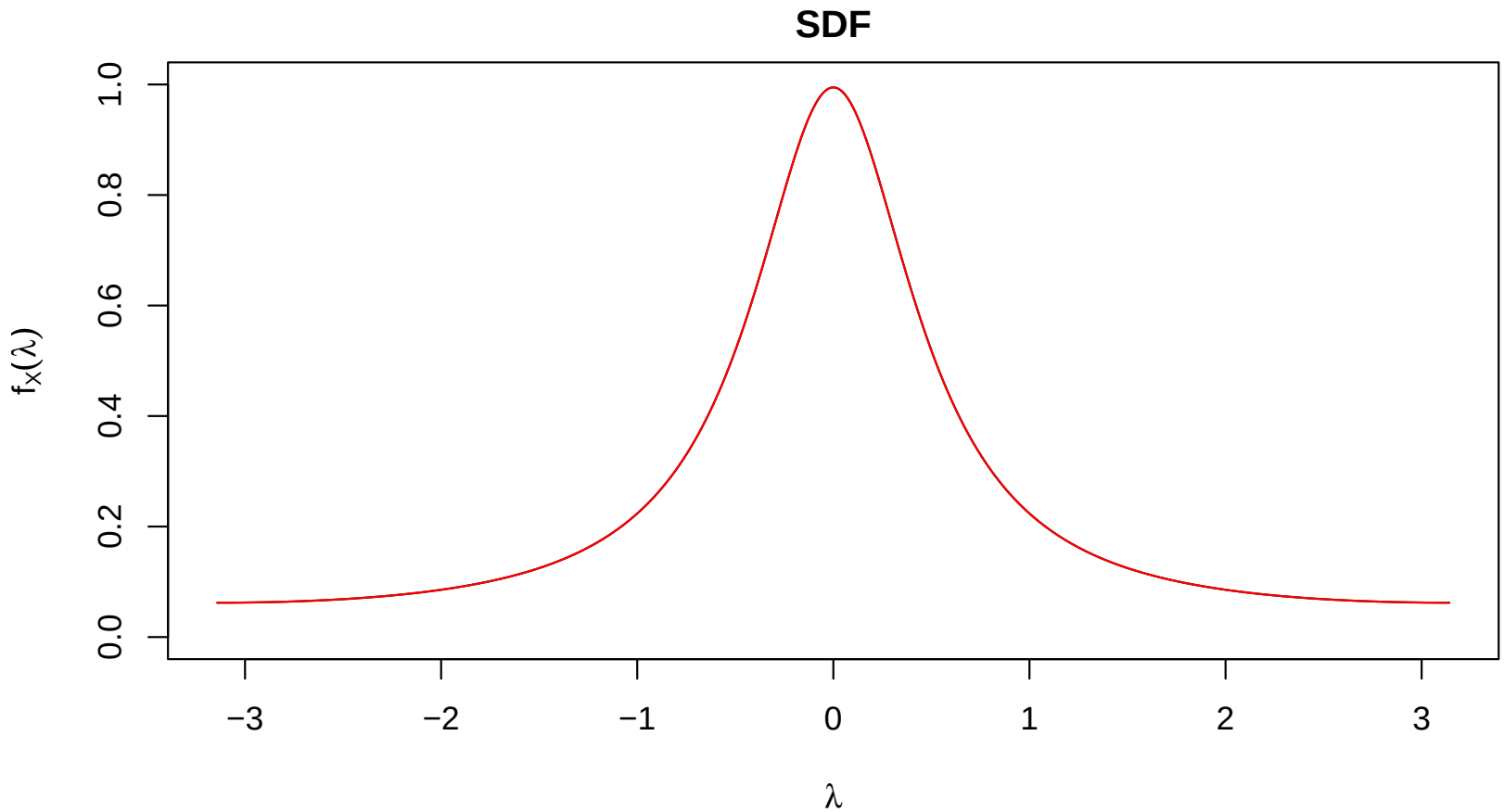
The Spectral Density Function

```
phi<-0.6
hvec<-0:30
acvf.vals<-1/(1-phi^2)*phi^hvec
lamvec<-seq(-pi,pi,by=0.001)
sdf.vec<-lamvec*0
hvec<-0:30
for(j in 1:length(lamvec)){
    sdf.vec[j]<-sum(exp(-1i*lamvec[j]*hvec[-1])*acvf.vals[2:31])+acvf.vals[1]+
                sum(exp(1i*lamvec[j]*hvec[-1])*acvf.vals[2:31])
}
sdf.vec<-sdf.vec/(2*pi)

exact.sdf<-1/(2*pi)*(1/Mod(1-phi*exp(-1i*lamvec))^2)
```

The Spectral Density Function

```
par(mar=c(4,4,2,2));plot(lamvec,sdf.vec,type='l',xlab=expression(lambda),
    ylab=expression(f[X](lambda)), ylim=range(0,1),main = 'SDF');lines(lamvec,exact.sdf,col='red')
```



The Spectral Density Function

Example: ARMA(1,1)

Consider the stationary ARMA$(1,1)$ model

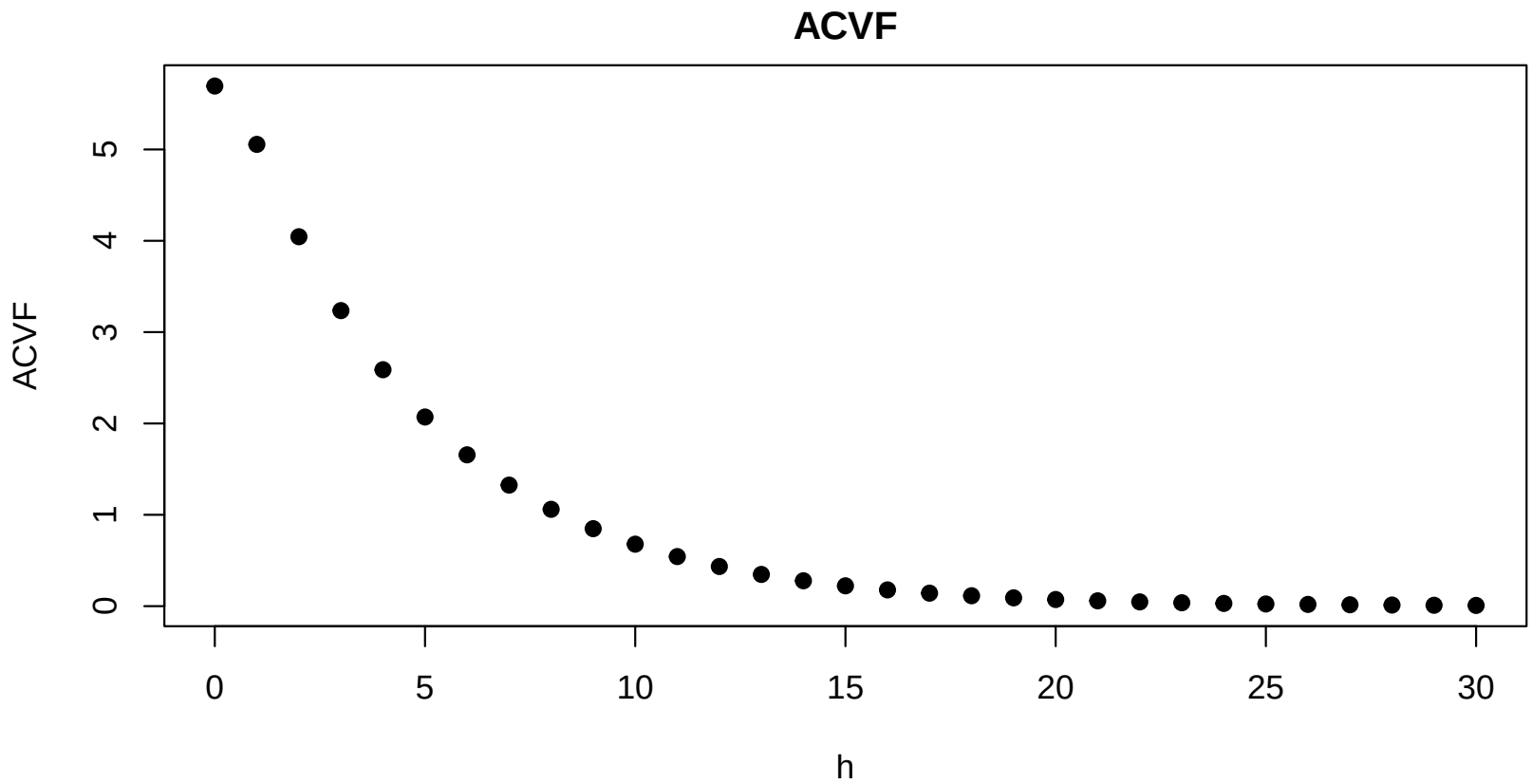
$$(1 - 0.8B)X_t = (1 + 0.5B)Z_t$$

We may inspect

- the ACVF
- the SDF

The Spectral Density Function

```
library(ltsa)
acvf.vals<-tacvfARMA(phi = 0.8, theta = -0.5, maxLag = 30)
par(mar=c(4,4,2,2));plot(0:30,acvf.vals,xlab='h', ylab='ACVF', main = 'ACVF',pch=19)
```

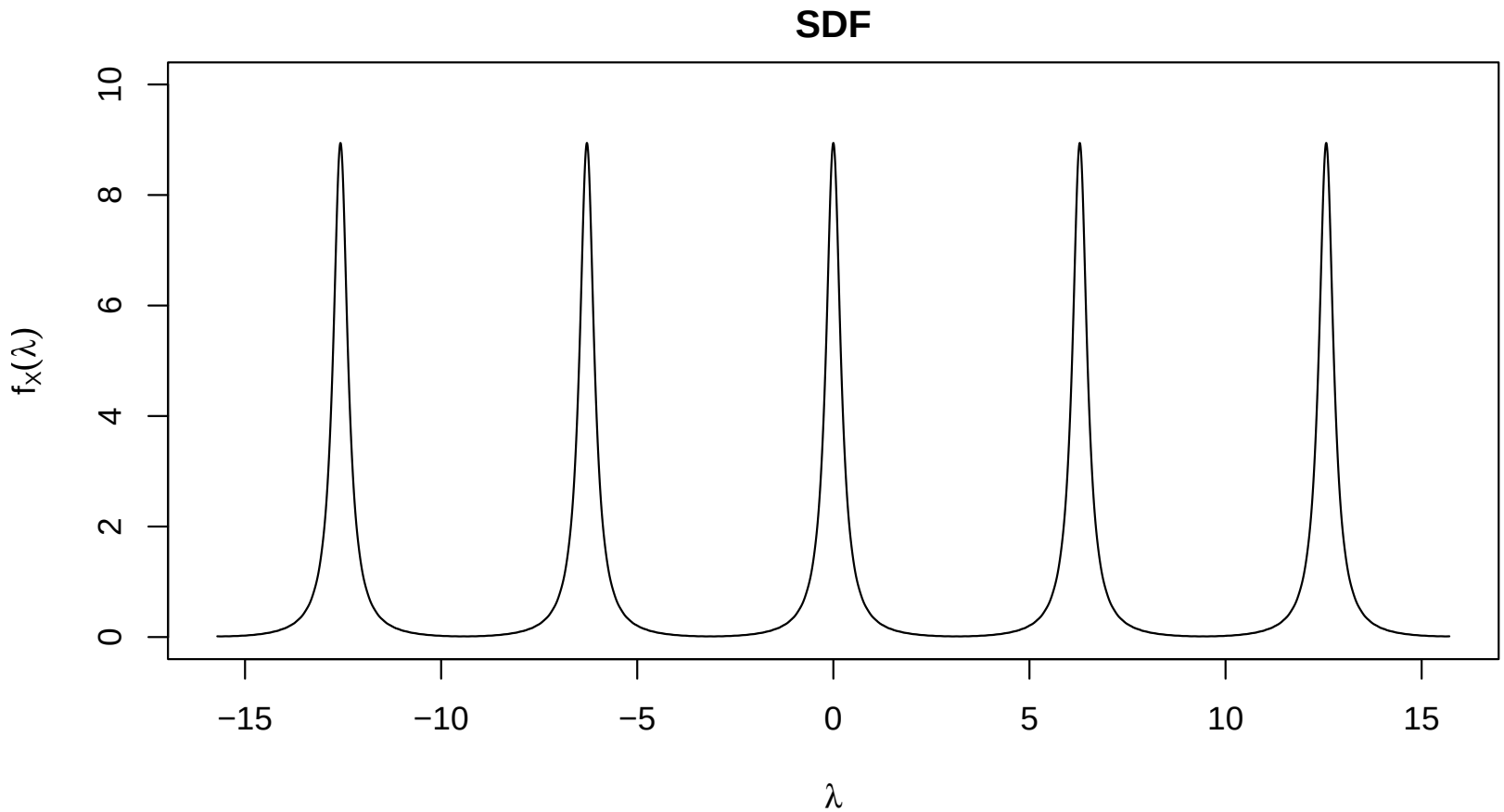


The Spectral Density Function

```
lamvec<-seq(-5*pi,5*pi,by=0.001)
sdf.vec<-lamvec*0
hvec<-0:30
for(j in 1:length(lamvec)){
    sdf.vec[j]<-sum(exp(-1i*lamvec[j]*hvec[-1])*acvf.vals[2:31])+acvf.vals[1]+
                sum(exp(1i*lamvec[j]*hvec[-1])*acvf.vals[2:31])
}
sdf.vec<-sdf.vec/(2*pi)
```

The Spectral Density Function

```
par(mar=c(4,4,2,2));plot(lamvec,sdf.vec,type='l',xlab=expression(lambda),
    ylab=expression(f[X](lambda)), ylim=range(0,10),main = 'SDF')
```



The Spectral Density Function

Example: ARMA(1,1)

It can be shown that

$$f_X(\lambda) = \frac{\sigma_Z^2}{2\pi} \left| \frac{1 + 0.5e^{-i\lambda}}{1 - 0.8e^{-i\lambda}} \right|^2$$

We plot on the range

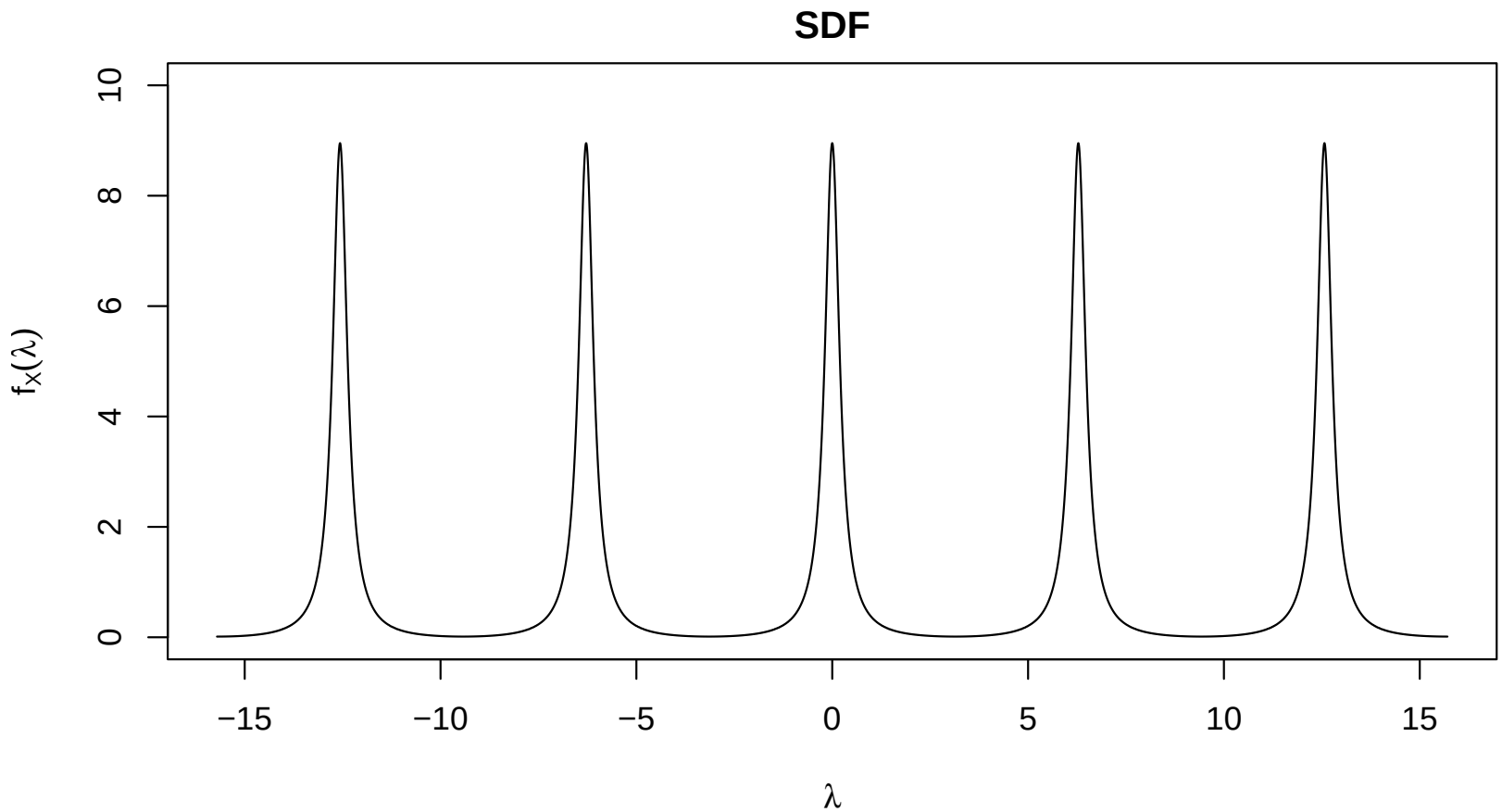
- $[-5\pi, 5\pi]$
- $[-\pi, \pi]$
- $[0, \pi]$

The Spectral Density Function

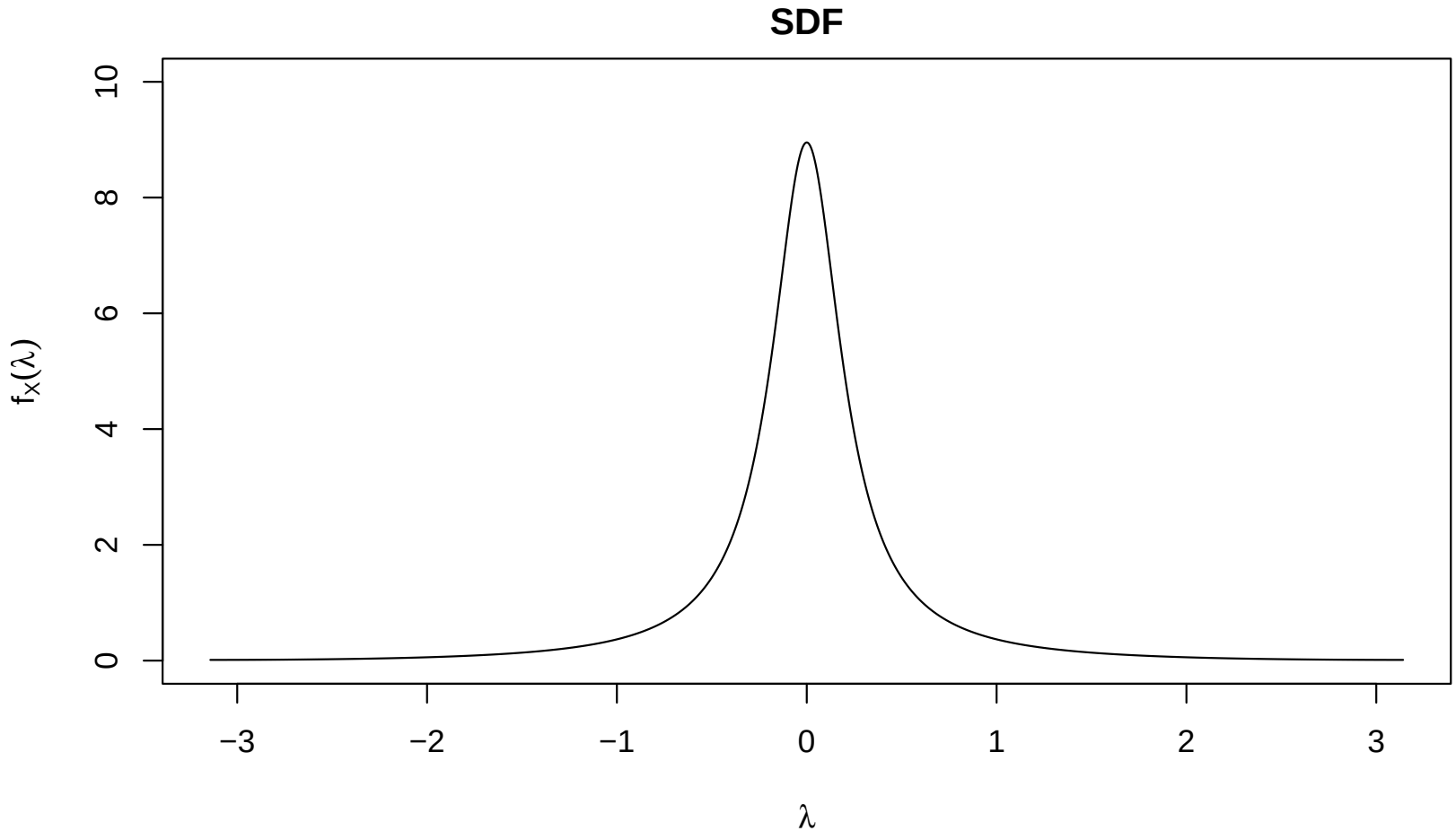
```
lamvec<-seq(-5*pi,5*pi,by=0.001)
sdf.vec<-(1/(2*pi))*Mod((1+0.5*exp(-1i*lamvec))/(1-0.8*exp(-1i*lamvec)))^2
```

The Spectral Density Function

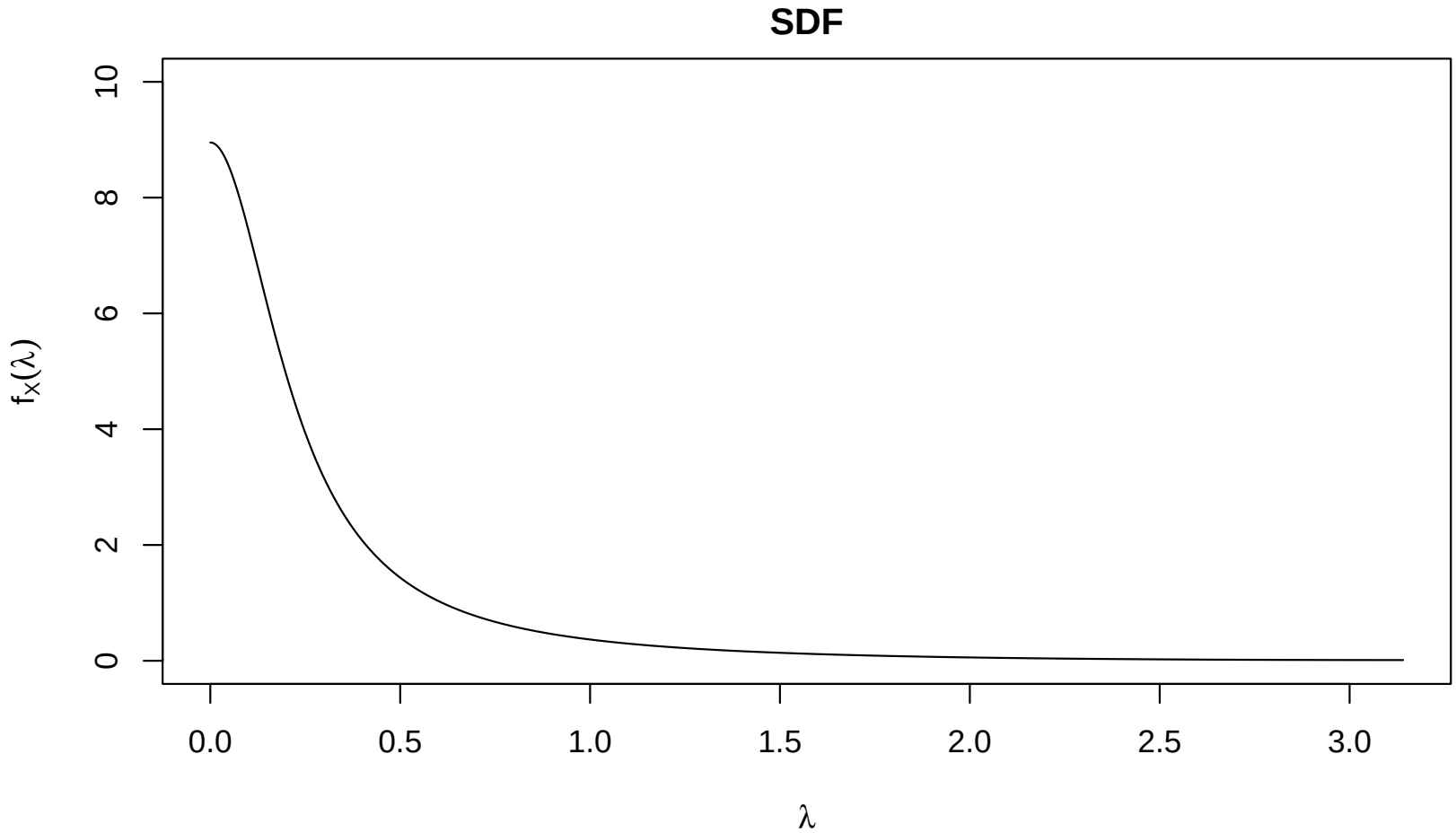
```
par(mar=c(4,4,2,2));plot(lamvec,sdf.vec,type='l',xlab=expression(lambda),
    ylab=expression(f[X](lambda)), ylim=range(0,10),main = 'SDF')
```



The Spectral Density Function

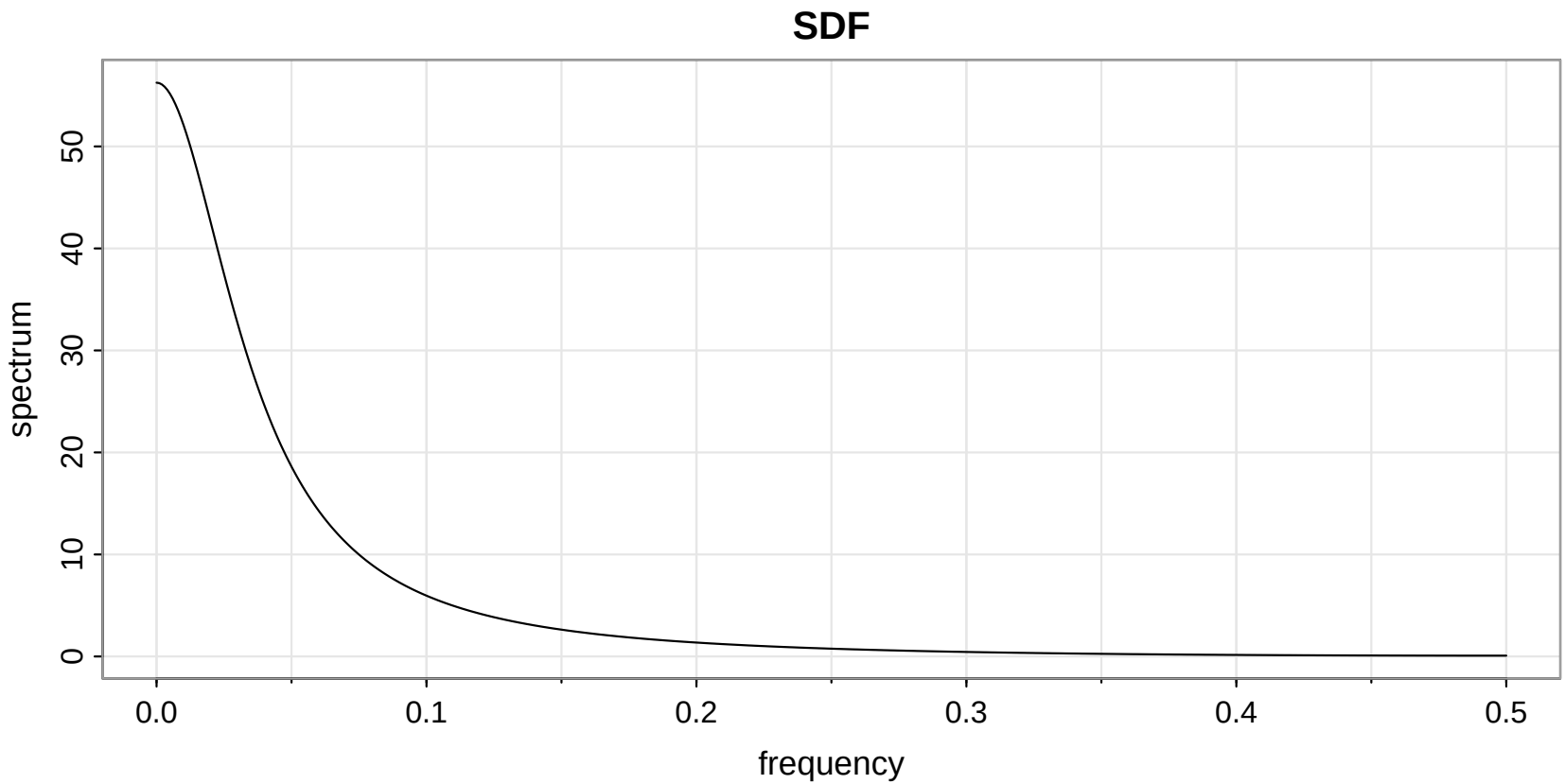


The Spectral Density Function



The Spectral Density Function

```
library(astsa)
par(mar=c(4,4,2,2))
arma.spec(ar = 0.8, ma = 0.5, var.noise = 1, n.freq = 500, main='SDF', frequency=1, plot=TRUE)
```



The Spectral Density Function

This computes the scaled version in terms of φ, where

$$\varphi = \frac{\lambda}{2\pi} \qquad \text{so that} \qquad -\frac{1}{2} < \varphi \leqslant \frac{1}{2}$$

in which case

$$f_X^*(\varphi) = \sum_{h=-\infty}^{\infty} e^{-i2\pi\varphi h} \gamma_X(h)$$

or more explicitly

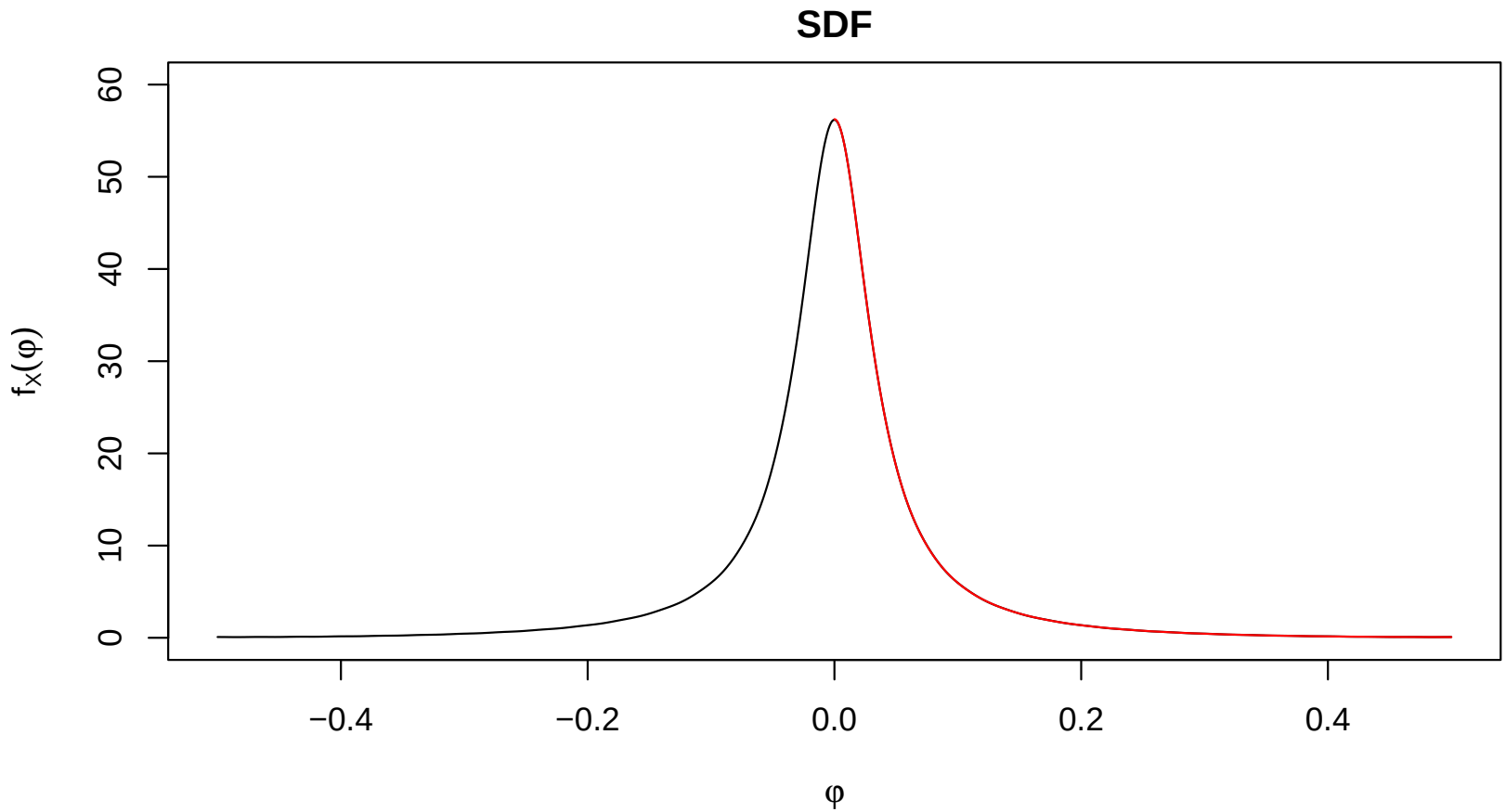
$$f_X^*(\varphi) = \sigma_Z^2 \left| \frac{1 + 0.5e^{-i2\pi\varphi}}{1 - 0.8e^{-i2\pi\varphi}} \right|^2$$

The Spectral Density Function

```
vvec0<-seq(-0.5,0.5,by=0.001)
sdf.vec0<-vvec0*0
hvec<-0:30
for(j in 1:length(vvec0)){
    sdf.vec0[j]<-sum(exp(-1i*2*pi*vvec0[j]*hvec[-1])*acvf.vals[2:31])+acvf.vals[1]+
                sum(exp(1i*2*pi*vvec0[j]*hvec[-1])*acvf.vals[2:31])
}
vvec1<-seq(0,0.5,by=0.001)
sdf.vec1<-Mod((1+0.5*exp(-1i*2*pi*vvec1))/(1-0.8*exp(-1i*2*pi*vvec1)))^2
```

The Spectral Density Function

```
par(mar=c(4,4,2,2));plot(vvec0,sdf.vec0,type='l',xlab=expression(varphi),
ylab=expression(f[X](varphi)), ylim=range(0,60),main = 'SDF'); lines(vvec1,sdf.vec1,col='red')
```



The Spectral Density Function

We have that

(i) $f_X(\lambda) = f_X(-\lambda)$ and

$$f_X(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \gamma_X(h)$$

$$= \frac{1}{2\pi} \left[\gamma_X(0) + \sum_{h=1}^{\infty} (e^{-i\lambda h} + e^{i\lambda h}) \gamma_X(h) \right] \quad \gamma(-h) = \gamma(h)$$

$$= \frac{1}{2\pi} \left[\gamma_X(0) + \sum_{h=1}^{\infty} 2\cos(\lambda h) \gamma_X(h) \right]$$

$$= \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \cos(\lambda h) \gamma_X(h) \qquad \text{as } \cos(-\lambda h) = \cos(\lambda h)$$

(ii) $f_X(\lambda) \geqslant 0, -\pi < \lambda \leqslant \pi.$

The Spectral Density Function

(iii) Inversion formula:

$$\begin{aligned}\int_{-\pi}^{\pi} e^{ih\lambda} f_X(\lambda)\, d\lambda &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \sum_{k=-\infty}^{\infty} e^{i\lambda(h-k)} \gamma_X(k) \right\} d\lambda \\ &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_X(k) \left\{ \int_{-\pi}^{\pi} e^{i\lambda(h-k)}\, d\lambda \right\} \\ &= \gamma_X(h).\end{aligned}$$

The Spectral Density Function

The last result follows as if $k \neq h$, setting $j = h - k$

$$\begin{aligned}\int_{-\pi}^{\pi} e^{i\lambda(h-k)}\, d\lambda &= \int_{-\pi}^{\pi} [\cos(\lambda j) + i \sin(\lambda j)]\, d\lambda \\ &= \int_{-\pi}^{\pi} \cos(\lambda j)\, d\lambda + i \int_{-\pi}^{\pi} \sin(\lambda j)\, d\lambda \\ &= 0.\end{aligned}$$

The Spectral Density Function

Note: The summability condition

$$\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty$$

is not a necessary condition for the existence of a SDF.

The Spectral Density Function

There is a 1-1 correspondence between ACVFs and spectral density functions; if

(i) $f_X(\lambda) \geqslant 0$, for all $\lambda \in (0, \pi]$.

(ii) For $h = 0, \pm 1, \pm 2, \ldots$, if

$$\gamma_X(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f_X(\lambda)\, d\lambda$$

then f_X is the SDF of $\{X_t\}$ with ACVF $\gamma_X(h)$. Note that

$$\gamma_X(h) = \int_{-\pi}^{\pi} [\cos(h\lambda) + i \sin(h\lambda)] f_X(\lambda) d\lambda = \int_{-\pi}^{\pi} \cos(h\lambda) f_X(\lambda) d\lambda.$$

The Spectral Density Function

A real-valued function f defined on $(-\pi, \pi]$ is the SDF of a stationary process if and only if

(i) f is even: $f(\lambda) = f(-\lambda)$ for all λ.

(ii) $f(\lambda) > 0$

(iii)

$$\int_{-\pi}^{\pi} f(\lambda)\, d\lambda < \infty$$

The Spectral Density Function

Suppose f satisfies (i)-(iii), then

$$\gamma_X(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f_X(\lambda)\, d\lambda$$

is even. Let $\mathbf{a} \in \mathbb{R}^n$. Then

$$\mathbf{a}^\top \mathbf{\Gamma} \mathbf{a} = \sum_{r=1}^{n}\sum_{s=1}^{n} a_r \gamma_X(r-s) a_s = \int_{-\pi}^{\pi} \sum_{r=1}^{n}\sum_{s=1}^{n} a_r a_s e^{i\lambda(r-s)} f(\lambda)\, d\lambda$$

$$= \int_{-\pi}^{\pi} \left| \sum_{r=1}^{n} a_r e^{i\lambda r} \right|^2 f(\lambda)\, d\lambda \geqslant 0$$

so $\gamma_X(.)$ is non-negative definite, and hence the ACVF for a stationary process.

The Spectral Density Function

Note:

- f_X is the Fourier transform of γ_X
- γ_X is the inverse Fourier transform of f_X.

The properties of the SDF f_X are similar to a pdf, with the additional condition that f_X must be symmetric about zero, and with the integrability relaxed to

$$\int_{-\pi}^{\pi} f_X(\lambda)\, d\lambda = C < \infty$$

rather than having $C = 1$.

Spectral Representation Theorem

Theorem: Uniqueness

A function $\gamma(.)$, $\gamma : \mathbb{Z} \longrightarrow \mathbb{R}$, is an autocovariance function if and only if there exists a right continuous, non-decreasing bounded function F on $[-\pi, \pi]$ with $F(-\pi) = 0$, and for all $h \in \mathbb{Z}$

$$\gamma(h) = \int_{(-\pi,\pi]} e^{ih\lambda} \, dF(\lambda).$$

As the corresponding series $\{\gamma(h)\}$ is real-valued, F satisfies

$$\int_{(a,b]} dF(x) = \int_{(-b,-a]} dF(x)$$

Spectral Representation Theorem

F is a generalized distribution function as $F(\lambda)/F(\pi)$ is a distribution function on $[-\pi, \pi]$.

F is termed the spectral distribution function; it can be purely discrete, absolutely continuous, or mixed. If F is absolutely continuous

$$F(\lambda) = \int_{-\pi}^{\lambda} f(x)\, dx \qquad \lambda \in (-\pi, \pi]$$

then f is the spectral density function.

Example

Example: Harmonic process

Suppose

$$X_t = U\cos(\omega t) + V\sin(\omega t)$$

where U and V are uncorrelated, with

$$\mathbb{E}[U] = \mathbb{E}[V] = 0 \qquad \text{Var}[U] = \text{Var}[V] = 1$$

Then by a previous result

$$\gamma_X(h) = \cos(\omega h) \qquad h \in \mathbb{Z}$$

Example

Example: Harmonic process

Now

$$\sum_{h=-\infty}^{\infty} |\gamma_X(h)| = \sum_{h=-\infty}^{\infty} |\cos(\omega h)| = \infty$$

However, let

$$F_X(\lambda) = \begin{cases} 0 & \lambda < -\omega \\ \frac{1}{2} & -\omega \leqslant \lambda < \omega \\ 1 & \omega \leqslant \lambda \end{cases}$$

Example

Example: Harmonic process

Then

$$f_X(\lambda) = \begin{cases} \frac{1}{2} & \lambda = -\omega \\ \frac{1}{2} & \lambda = \omega \\ 0 & \text{otherwise} \end{cases}$$

and

$$\int_{(-\pi,\pi)} e^{ih\lambda}\, dF_X(\lambda) = \frac{1}{2}e^{-ih\omega} + \frac{1}{2}e^{ih\omega} = \cos(h\omega)$$

Example: Extension

Example: Superposition of harmonic processes

Suppose

$$X_t = \sum_{j=1}^{k} [U_j \cos(\omega_j t) + V_j \sin(\omega_j t)]$$

where $U_1, \ldots, U_k$ and $V_1, \ldots, V_k$ are uncorrelated, with

$$\mathbb{E}[U_j] = \mathbb{E}[V_j] = 0 \qquad \text{Var}[U_j] = \text{Var}[V_j] = \sigma_j^2$$

and where $0 < \omega_1 < \cdots < \omega_k < \pi$. Then

$$\gamma_X(h) = \sum_{j=1}^{k} \sigma_j^2 \cos(\omega_j h) \qquad h \in \mathbb{Z}.$$

Example: Extension

Example: Superposition of harmonic processes

Then

$$F_X(\lambda) = \sum_{j=1}^{k} \sigma_j^2 F_j(\lambda)$$

where

$$F_j(\lambda) = \begin{cases} 0 & \lambda < -\omega_j \\ \frac{1}{2} & -\omega_j \leqslant \lambda < \omega_j \\ 1 & \omega_j \leqslant \lambda \end{cases}$$

Note that

$$\mathrm{Var}[X_t] = \gamma_X(0) = \sum_{j=1}^{k} \sigma_j^2$$

Spectral decomposition

This result shows that the variance of $\{X_t\}$ can be decomposed into the variances contributed by the harmonic components at different frequencies

$$\omega_1, \ldots, \omega_k$$

In the limit as $k \longrightarrow \infty$

$$X_t \overset{\text{m.s.}}{=} \int_{(-\pi,\pi]} e^{it\lambda} dZ(\lambda)$$

where $\{Z(\lambda) : -\pi < \lambda \leqslant \pi\}$ is a non-decreasing stochastic process.

In general, Z is a complex-valued process which has uncorrelated increments.

Spectral decomposition

A complex-valued process, $\{X_t\}$, is a stochastic process with complex-valued elements

$$X_t = X_{1,t} + iX_{2,t}$$

say, where

$$\begin{aligned}
\mathbb{E}[X_t] &= \mathbb{E}[X_{1,t}] + i\mathbb{E}[X_{2,t}] \\
\mathbb{E}[X_t X_{t+h}] &= \mathbb{E}[(X_{1,t} + iX_{2,t})(X_{1,t+h} + iX_{2,t+h}] \\
&= \mathbb{E}[X_{1,t}X_{1,t+h} - X_{2,t}X_{2,t+h}] \\
&\quad + i\mathbb{E}[X_{1,t}X_{2,t+h} + X_{2,t}X_{1,t+h}]
\end{aligned}$$

etc.

Spectral decomposition

We have

$$X_t = \int_{(-\pi,\pi]} e^{it\lambda} dZ(\lambda)$$

$$= \int_{(-\pi,\pi]} \left[\cos(t\lambda) + i\sin(t\lambda)\right]\ dZ(\lambda)$$

where $Z(\lambda)$ is a random process on $(-\pi, \pi]$. By symmetry, we need only consider $Z(\lambda)$ for $0 \leqslant \lambda \leqslant \pi$.

Spectral decomposition

For finite k,

$$dZ(\lambda) = \begin{cases} \dfrac{U_j + iV_j}{2} & \lambda = -\omega_j,\ j = 1, \ldots, k \\ \dfrac{U_j - iV_j}{2} & \lambda = \omega_j,\ j = 1, \ldots, k \\ 0 & \text{otherwise} \end{cases}$$

Here

$$\mathbb{E}[dZ(\lambda)\overline{dZ(\lambda)}] = \mathbb{E}[(U_j^2 + V_j^2)/4] = \frac{\sigma_j^2}{2}$$

for $\lambda = \pm\omega_j$, $j = 1, \ldots, k$ and zero otherwise.

Example: White noise process

Example: $Z_t \sim \text{WN}(0, \sigma_Z^2)$

If $Z_t \sim \text{WN}(0, \sigma_Z^2)$, then $\gamma_Z(0) = \sigma_Z^2$, $\gamma_Z(h) = 0$ for all $h \neq 0$. Then, for $-\pi < \lambda \leqslant \pi$,

$$\begin{aligned} f_Z(\lambda) &= \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma_Z(h) \\ &= \frac{1}{2\pi} \left[\cos(0) - i \sin(0)\right] \sigma_Z^2 \\ &= \frac{\sigma_Z^2}{2\pi} \end{aligned}$$

Spectral decomposition

λ is an angular frequency:

- For a time series with dominant high frequency components, $f_X(\lambda)$ is large when λ is near π, and the series exhibits rapid variation.
- For a time series with dominant low frequency components, $f_X(\lambda)$ is large when λ is near zero, and the series exhibits slow variation.

Spectral decomposition

Often we use the scaled frequency φ,

$$\varphi = \frac{\lambda}{2\pi} \qquad \text{so that} \qquad -\frac{1}{2} < \varphi \leqslant \frac{1}{2}$$

so that

$$f_X^*(\varphi) = \sum_{h=-\infty}^{\infty} e^{-i2\pi\varphi h} \gamma_X(h)$$

and

$$\gamma_X(h) = \int_{-1/2}^{1/2} e^{i2\pi\varphi h} f_X^*(\varphi)\, d\varphi$$

6.3 Representation of ARMA processes

Suppose $\Psi \equiv \{\psi_j : j = 0, \pm 1, \pm 2, \ldots\}$ is a linear filter with

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty.$$

and suppose $\{X_t\}$ is zero mean with SDF $f_X(.)$.

Let

$$\Psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j \qquad z \in \mathbb{C}.$$

and note that

$$\overline{\Psi(z)} = \Psi(\overline{z}).$$

6.3 Representation of ARMA processes

If

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j} = \Psi(B) X_t.$$

then $\{Y_t\}$ is stationary and zero mean, with

$$\gamma_Y(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_X(h + k - j)$$

by the previous result for ACVFs

6.3 Representation of ARMA processes

Then

$$
\begin{aligned}
f_Y(\lambda) &= \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \gamma_Y(h) \\
&= \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \left\{ \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_X(h+k-j) \right\} \\
&= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \left\{ \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\lambda h} \gamma_X(h+k-j) \right\} \\
&= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \left\{ \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} e^{-i\lambda (l+j-k)} \gamma_X(l) \right\}
\end{aligned}
$$

setting $l = h + k - j$.

6.3 Representation of ARMA processes

Therefore

$$f_Y(\lambda) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k e^{-i\lambda(j-k)} \left\{ \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} e^{-i\lambda l} \gamma_X(l) \right\}$$

$$= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k e^{-i\lambda(j-k)} f_X(\lambda)$$

$$= \left(\sum_{j=-\infty}^{\infty} \psi_j e^{-i\lambda j} \right) \left(\sum_{k=-\infty}^{\infty} \psi_k e^{i\lambda k} \right) f_X(\lambda)$$

$$= \Psi(e^{-i\lambda}) \Psi(e^{i\lambda}) f_X(\lambda)$$

$$\equiv \Psi(e^{-i\lambda}) \overline{\Psi(e^{-i\lambda})} f_X(\lambda)$$

6.3 Representation of ARMA processes

Therefore, the SDF of Y is

$$f_Y(\lambda) = |\Psi(e^{-i\lambda})|^2 f_X(\lambda) = \Psi(e^{-i\lambda})\Psi(e^{i\lambda}) f_X(\lambda)$$

- $\Psi(e^{-i\lambda})$ is the transfer function.
- $|\Psi(e^{-i\lambda})|^2$ is the power transfer function.

The transfer function defines the action of the filter on the process $\{X_t\}$.

Low Pass filter

Example: Low Pass filter

A low pass filter removes high frequency components. To remove frequency components beyond threshold λ_C, we would require a transfer function taking the value 1 for $-\lambda_C \leqslant \lambda \leqslant \lambda_C$, that is,

$$\Psi(e^{-i\lambda}) = \begin{cases} 1 & -\lambda_C \leqslant \lambda \leqslant \lambda_C \\ 0 & \text{otherwise} \end{cases}$$

We term this filter the ideal low pass filter.

Low Pass filter

Example: Low Pass filter

Let

$$\psi_j = \frac{1}{2\pi}\int_{-\lambda_C}^{\lambda_C} e^{ij\lambda}\, d\lambda = \begin{cases} \dfrac{\lambda_C}{\pi} & j = 0 \\ \dfrac{\sin(j\lambda_C)}{j\pi} & j \neq 0 \end{cases}$$

and define

$$\Psi_q(e^{-i\lambda}) = \sum_{j=-q}^{q} \psi_j e^{-ij\lambda}$$

for integer q. This filter is a finite Fourier series approximation to the ideal low pass filter.

Example: $X_t \sim \text{ARMA}(p, q)$

Suppose $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$, and define $\{X_t\}$ by

$$\Phi(B)X_t = \Theta(B)Z_t.$$

Thus

$$f_Z(\lambda) = \frac{\sigma_Z^2}{2\pi}$$

and as $X_t = \{\Phi(B)\}^{-1}\Theta(B)Z_t$, we have

$$f_X(\lambda) = \frac{\sigma_Z^2}{2\pi}\left|\frac{\Theta(e^{-i\lambda})}{\Phi(e^{-i\lambda})}\right|^2 = \frac{\sigma_Z^2}{2\pi}\frac{\left|\Theta(e^{-i\lambda})\right|^2}{\left|\Phi(e^{-i\lambda})\right|^2}$$

by the general result.

6.4 Model-free estimation

The SDF can be estimated from time-domain samples. Let $x_{1:n}$ be the time-domain data, and let

$$\omega_k = \frac{2\pi k}{n}$$

for k equal to

$$-\left\lfloor \frac{n-1}{2} \right\rfloor, \ldots, \left\lfloor \frac{n}{2} \right\rfloor$$

The set $\{\omega_k\}$ are termed the Fourier frequencies.

6.4 Model-free estimation

For each k, define the $n \times 1$ vector $\boldsymbol{e}_k$ by

$$\boldsymbol{e}_k = \frac{1}{\sqrt{n}}\left(e^{i\omega_k}, e^{2i\omega_k}, \ldots, e^{ni\omega_k}\right)^\top$$

so that

$$\overline{\boldsymbol{e}}_j^\top \boldsymbol{e}_k = 0$$

unless $j = k$, in which case $\overline{\boldsymbol{e}}_k^\top \boldsymbol{e}_k = 1$.

6.4 Model-free estimation

$\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ form an orthonormal basis for $\mathbb{C}^n$, that is, for any n-dimensional complex number $\mathbf{x}$, we have

$$\mathbf{x} = \sum_{k=-\lfloor (n-1)/2 \rfloor}^{\lfloor n/2 \rfloor} a_k \boldsymbol{e}_k$$

where

$$a_k = \overline{\boldsymbol{e}}_k^\top \mathbf{x} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t e^{-it\omega_k}$$

The vector $(a_1, \ldots, a_n)^\top$ is the discrete Fourier transform of $\mathbf{x}$.

6.4 Model-free estimation

Note that

$$x_t = \sum_{k=-\lfloor (n-1)/2 \rfloor}^{\lfloor n/2 \rfloor} a_k \left[\cos(\omega_k t) + i \sin(\omega_k t) \right]$$

is the discrete version of the spectral representation theorem.

The Periodogram

The periodogram, $\widehat{I}_n(\lambda)$ is defined by

$$\widehat{I}_n(\lambda) = \frac{1}{n}\left|\sum_{t=1}^{n} x_t e^{-it\lambda}\right|^2 \qquad -\pi < \lambda \leqslant \pi.$$

Note that $\widehat{I}_n(\omega_k) = |a_k|^2$, and

$$\bar{\mathbf{x}}^\top \mathbf{x} = \sum_{k=-\lfloor (n-1)/2 \rfloor}^{\lfloor n/2 \rfloor} |a_k|^2 = \sum_{k=-\lfloor (n-1)/2 \rfloor}^{\lfloor n/2 \rfloor} \widehat{I}_n(\omega_k)$$

which gives a decomposition of the total variation in $\mathbf{x}$ via components at the ω_k.

The Periodogram

Note that

$$\begin{aligned}
\widehat{I}_n(\omega_k) &= \left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t e^{-it\omega_k}\right)\left(\frac{1}{\sqrt{n}}\sum_{s=1}^{n} x_s e^{is\omega_k}\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n} x_t x_s e^{-i(t-s)\omega_k} \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n} (x_t-\overline{x})(x_s-\overline{x}) e^{-i(t-s)\omega_k} \\
&= \sum_{h=-(n-1)}^{n-1} \widehat{\gamma}_n(h) e^{-ih\omega_k} \qquad (h = s-t)
\end{aligned}$$

as $\sum_{t=1}^{n} e^{-it\omega_k} = 0$ if $\omega_k \neq 0$.

The Periodogram

Thus $\widehat{I}_n(\lambda)$ is a sample-based estimate of $2\pi f_X(\lambda)$; it is **not** consistent, however. For $0 < \lambda_1 < \ldots < \lambda_m < \pi$,

$$(\widehat{I}_n(\lambda_1), \ldots, \widehat{I}_n(\lambda_m))^\top \xrightarrow{d} (V_1, \ldots, V_n)^\top$$

as $n \longrightarrow \infty$, where

$$V_j \sim \mathit{Exponential}(1/(2\pi f_X(\lambda_j))).$$

That is, as $n \longrightarrow \infty$

$$\mathbb{E}[\widehat{I}_n(\lambda_j)] \longrightarrow 2\pi f_X(\lambda_j)$$

so $\widehat{I}_n(\lambda_j)$ is asymptotically unbiased, but $\widehat{I}_n(\lambda_j)$ converges in distribution, not in probability or almost surely.

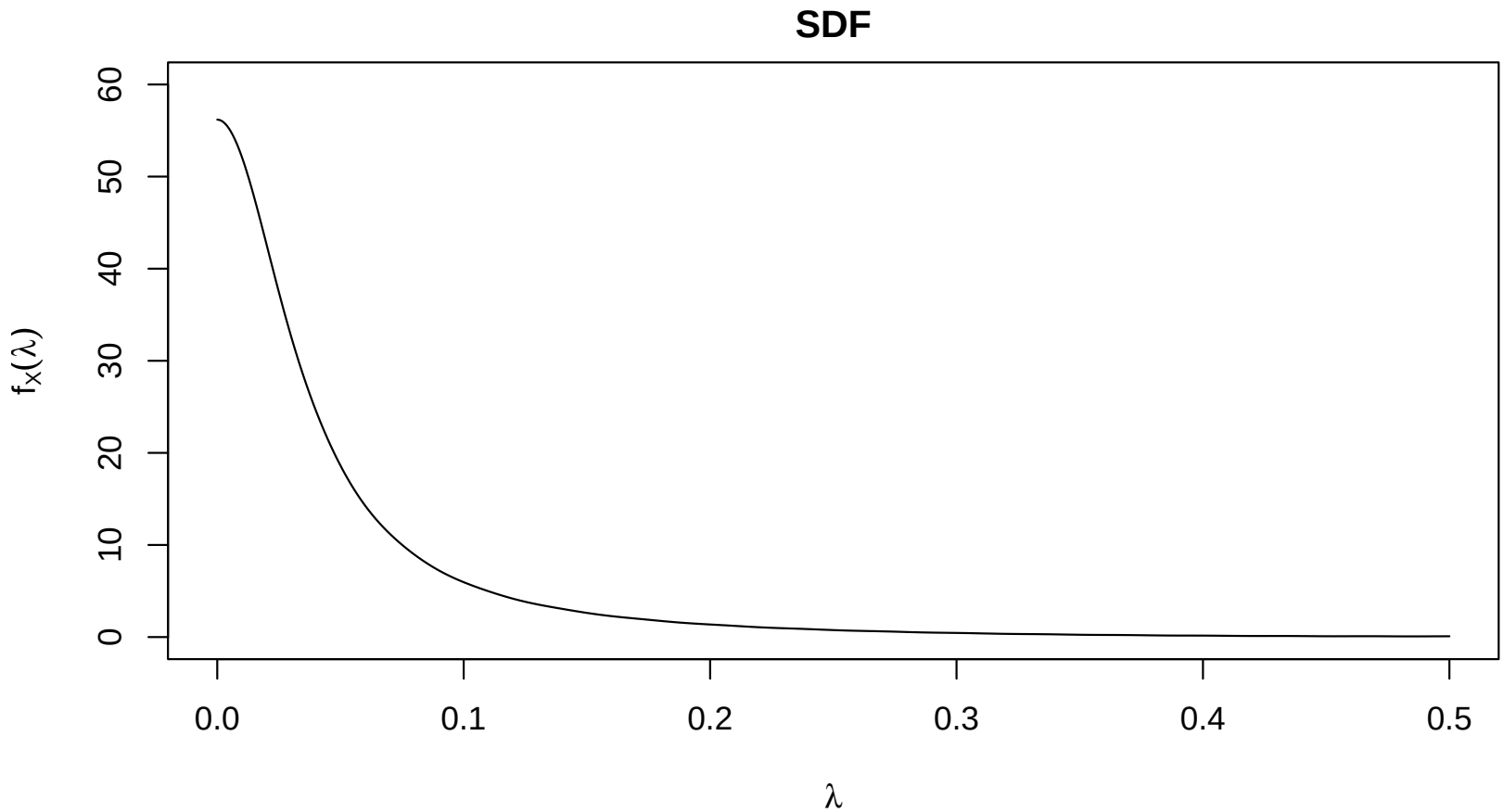
The Periodogram

Example: ARMA(1,1)

```
acvf.vals<-tacvfARMA(phi = 0.8, theta = -0.5, maxLag = 30)
lamvec<-seq(0,0.5,by=0.001)
sdf.vec<-lamvec*0
hvec<-0:30
for(j in 1:length(lamvec)){
    sdf.vec[j]<-sum(exp(-1i*2*pi*lamvec[j]*hvec[-1])*acvf.vals[2:31])+acvf.vals[1]+
                sum(exp(1i*2*pi*lamvec[j]*hvec[-1])*acvf.vals[2:31])
}
```

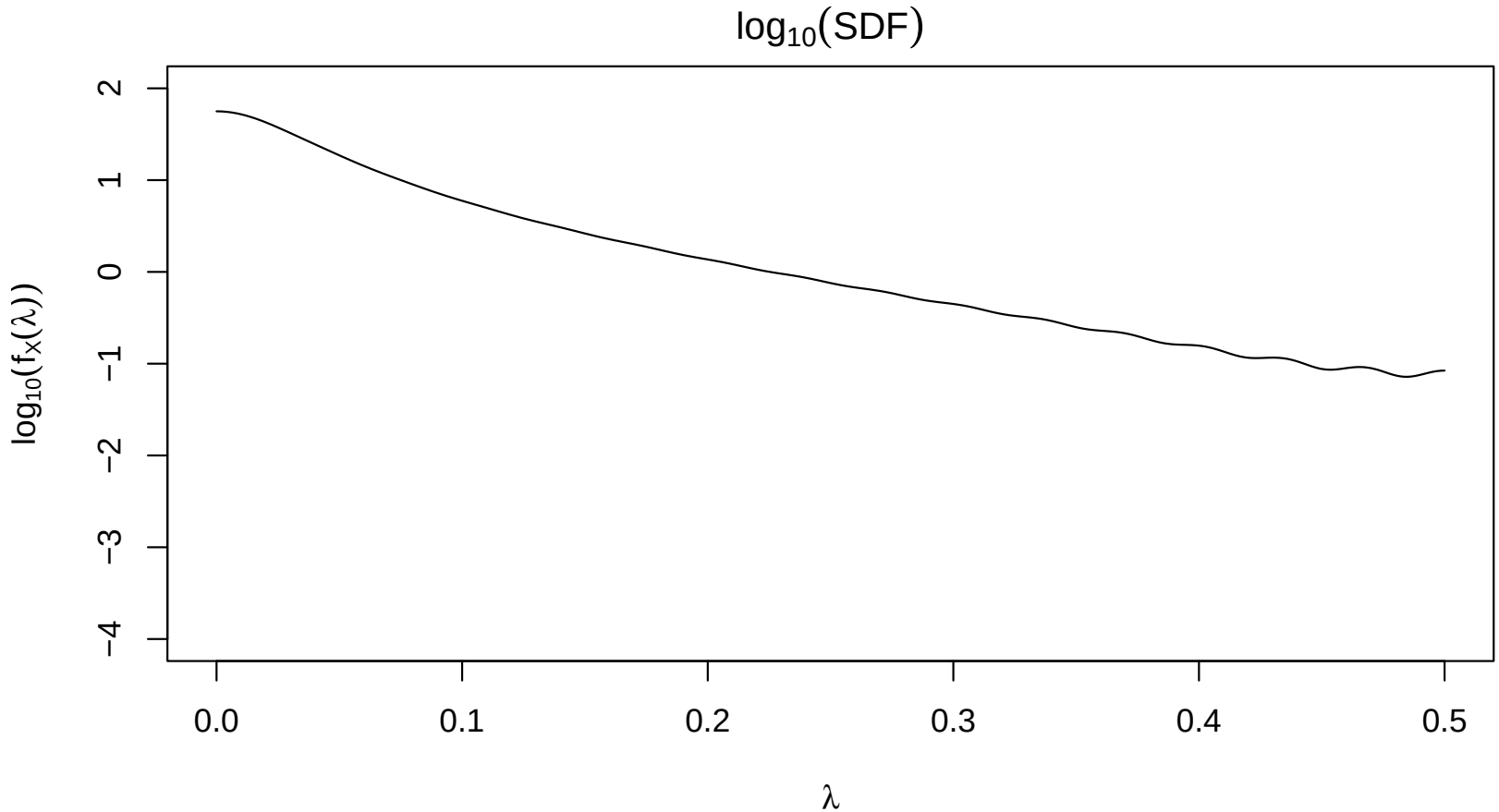
The Periodogram

```
par(mar=c(4,4,2,2));plot(lamvec,sdf.vec,type='l',xlab=expression(lambda),
    ylab=expression(f[X](lambda)), ylim=range(0,60),main = 'SDF')
```



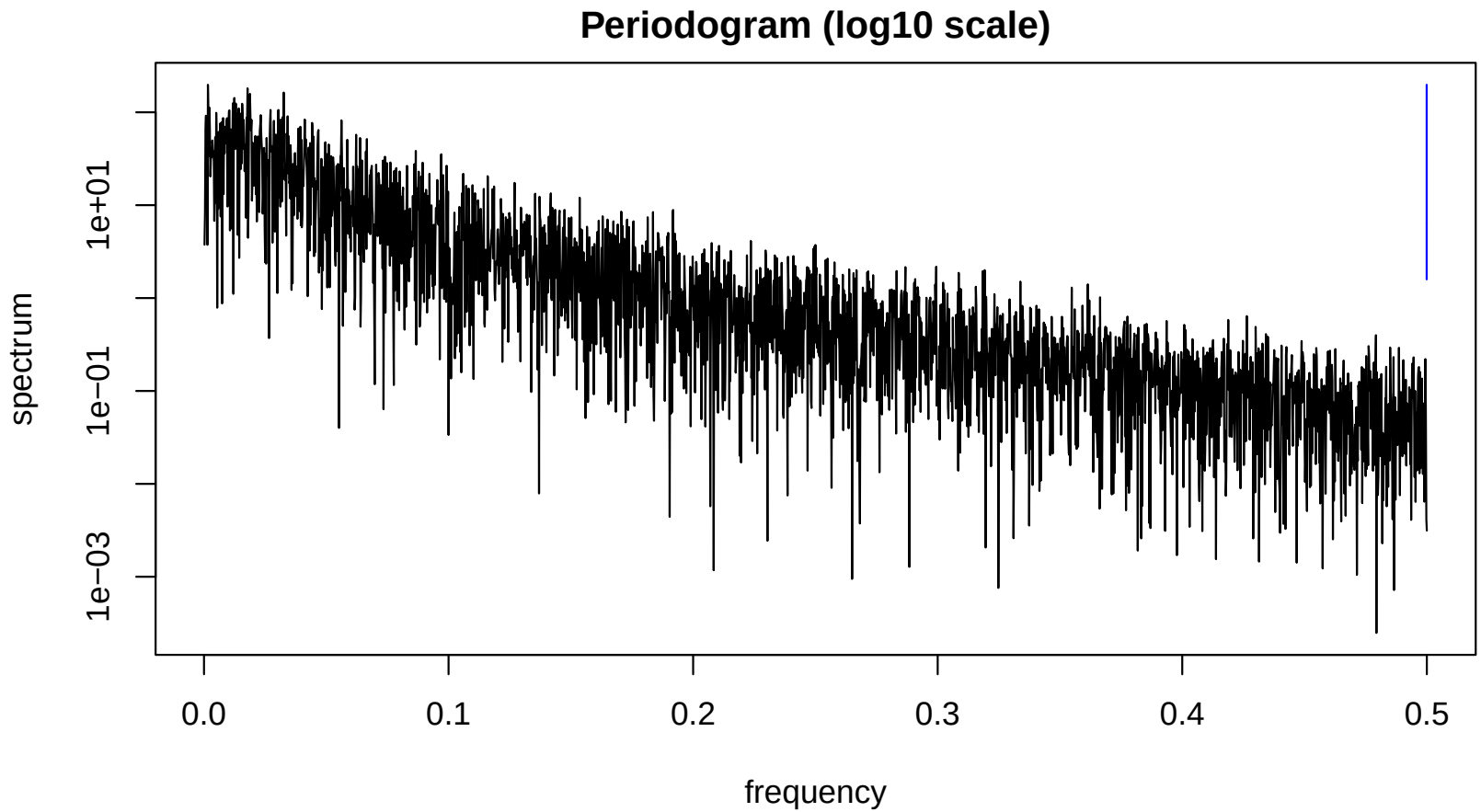
The Periodogram

```
par(mar=c(4,4,2,2));plot(lamvec,log10(sdf.vec),type='l',xlab=expression(lambda),
    ylab=expression(log[10](f[X](lambda))), ylim=range(-4,2),main = expression(log[10](SDF)))
```



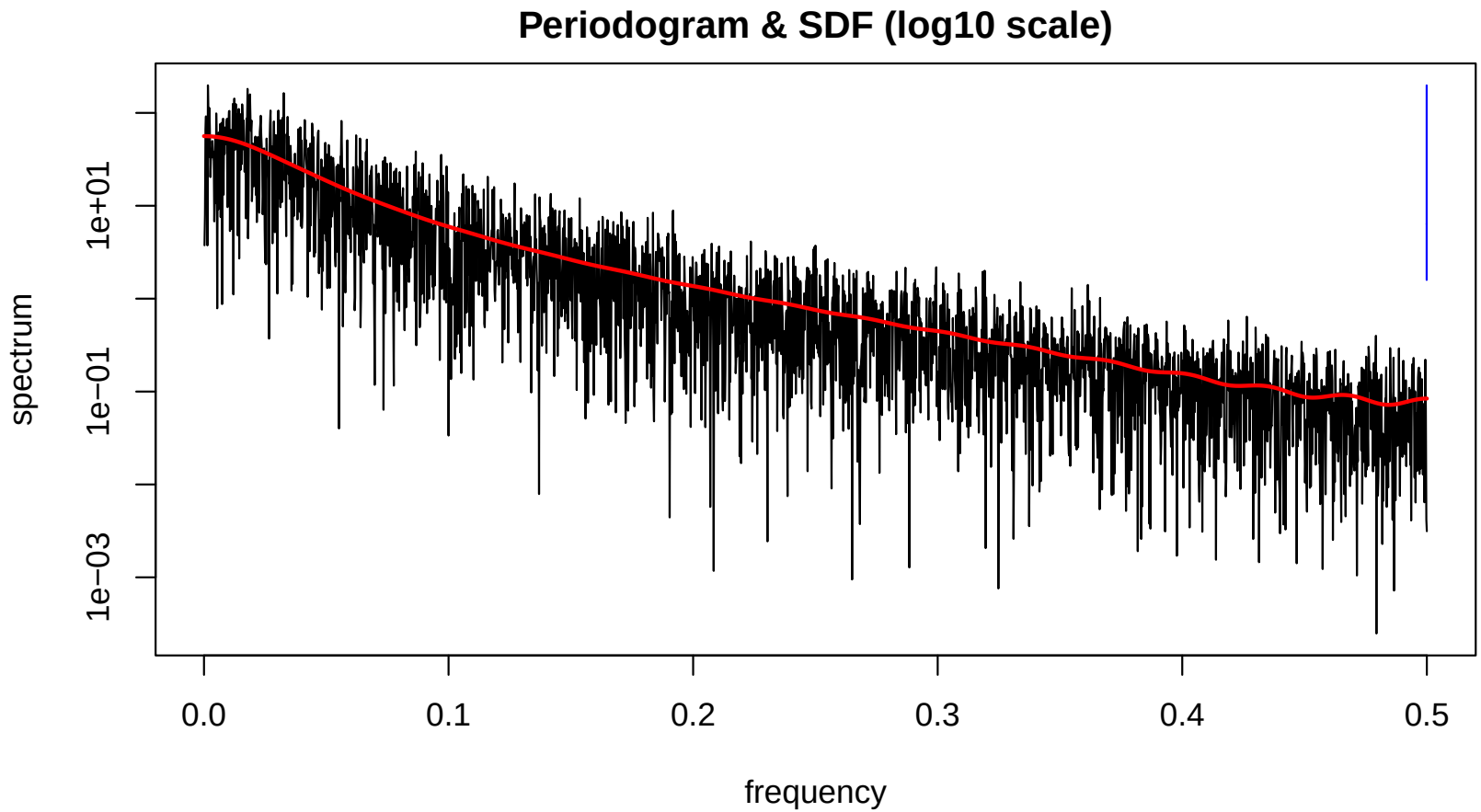
The Periodogram

```
set.seed(3); Xt<-arima.sim(n=5000,model=list(ar=0.8,ma=0.5))
par(mar=c(4,4,2,2));spectrum(Xt, main='Periodogram (log10 scale)', taper=0,detrend=FALSE)
```

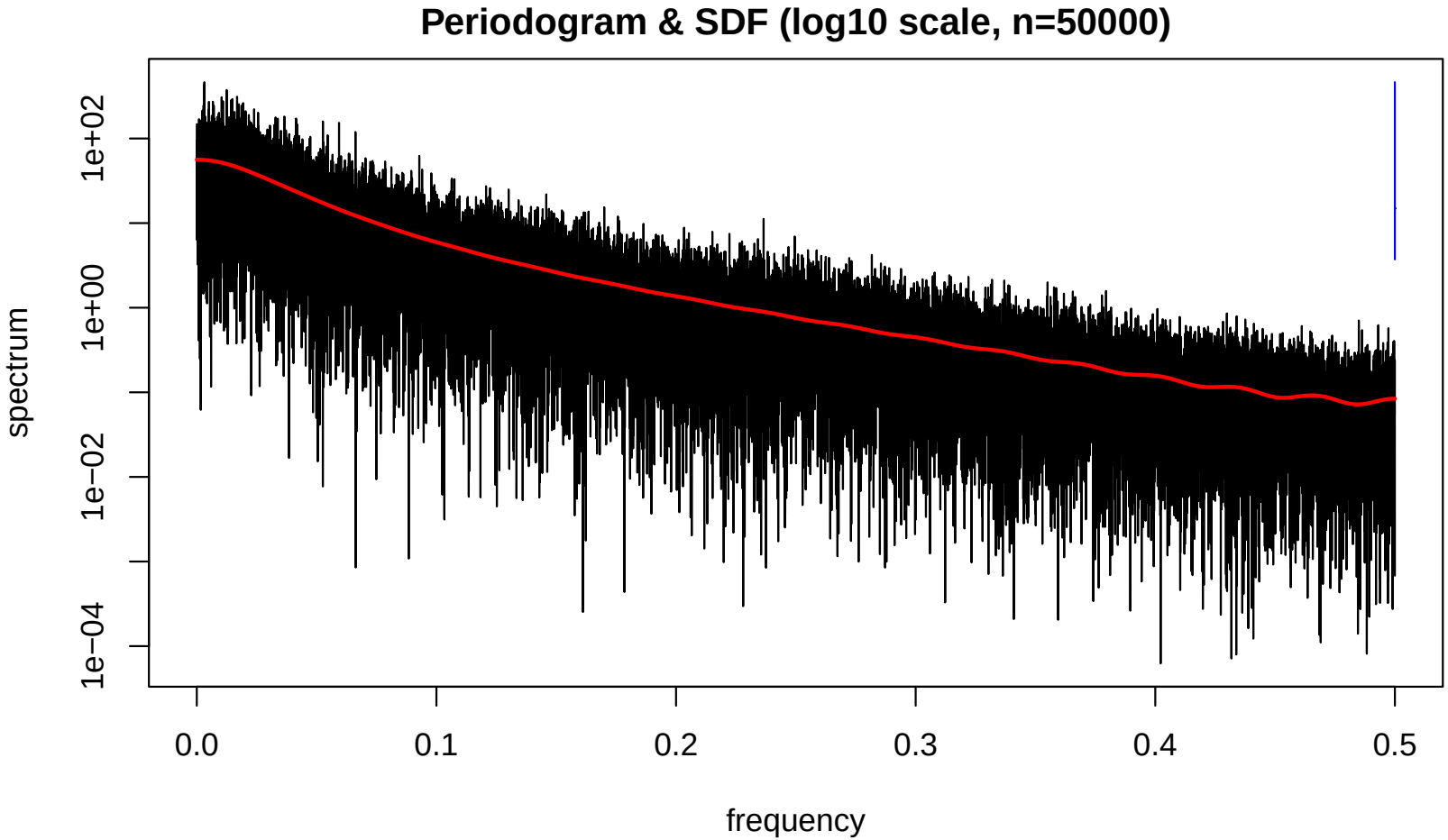


The Periodogram

```
par(mar=c(4,4,2,2));spectrum(Xt, main='Periodogram & SDF (log10 scale)', taper=0,detrend=FALSE)
lines(lamvec,sdf.vec,col='red',lwd=2)
```



The Periodogram

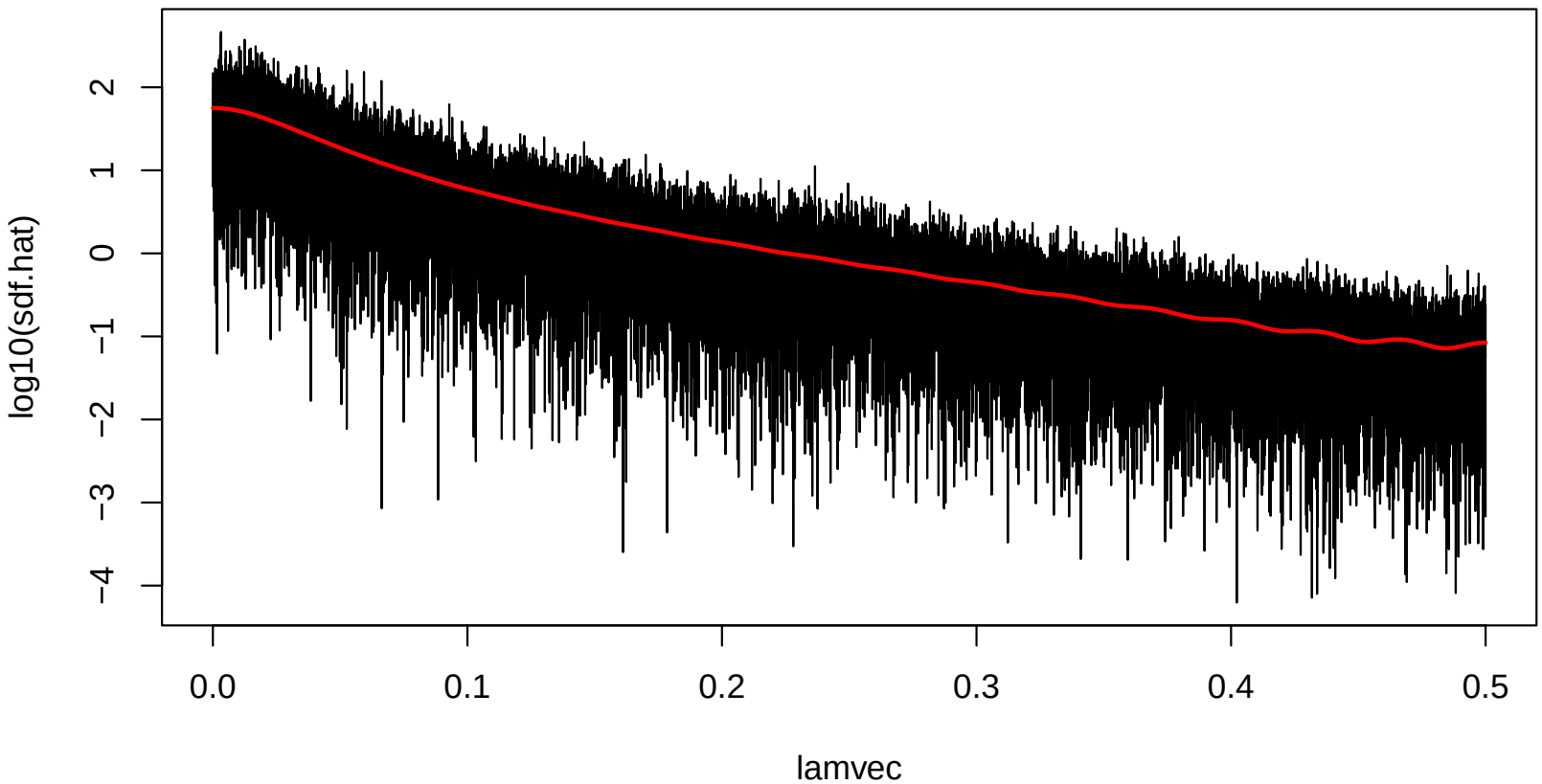


The Periodogram

```
sdf.hat<-spectrum(Xt, taper=0,detrend=FALSE,plot=FALSE)$spec
n0<-length(sdf.hat)
lamvec<-c(1:n0)/(2*n0)
sdf.vec<-lamvec*0
hvec<-0:30
for(j in 1:length(lamvec)){
    sdf.vec[j]<-sum(exp(-1i*2*pi*lamvec[j]*hvec[-1])*acvf.vals[2:31])+acvf.vals[1]+
                sum(exp(1i*2*pi*lamvec[j]*hvec[-1])*acvf.vals[2:31])
}
```

The Periodogram

```
par(mar=c(4,4,2,2));plot(lamvec,log10(sdf.hat),type='l')
lines(lamvec,log10(sdf.vec),col='red',lwd=2)
```



The Periodogram

If in the above notation,

$$V_j \sim Exponential(1/(2\pi f_X(\lambda_j))).$$

then

$$\frac{V_j}{2\pi f_X(\lambda_j)} \sim Exponential(1)$$

which we can verify in the sample.

Note that in the previous calculations, the factor of 2π is already incorporated to account for the output of `spectrum`.

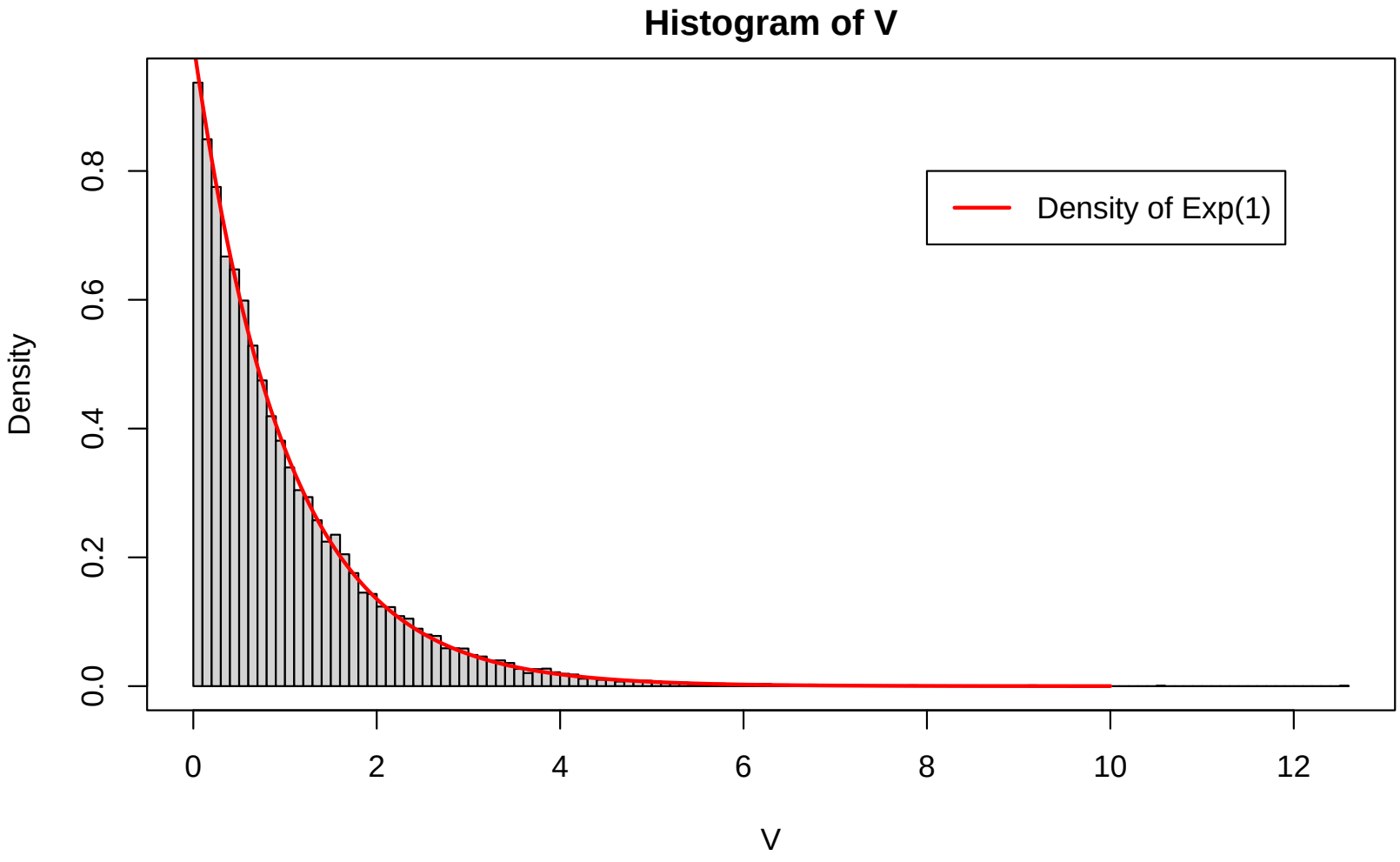
The Periodogram

```
Vvec<-as.numeric(sdf.hat/sdf.vec)
mean(Vvec); var(Vvec)

## [1] 1.0135
## [1] 1.025988

par(mar=c(4,4,2,2))
hist(Vvec,nclass=100, main='Histogram of V', xlab='V', freq=FALSE)
xv<-seq(0,10,by=0.001)
lines(xv,dexp(xv),col='red',lwd=2)
legend(8,0.8,c('Density of Exp(1)'),lwd=2,col='red')
box()
```

The Periodogram



The Periodogram

To obtain a consistent estimator, a local averaging process is used: define

$$\widehat{f}_n(\lambda) = \frac{1}{2\pi} \sum_{|j| \leqslant m_n} w_n(j) \widehat{I}_n(g(n,\lambda) + 2\pi j/n)$$

where bandwidths m_n satisfy $m_n \longrightarrow \infty$, $m_n/n \longrightarrow 0$ as $n \longrightarrow \infty$, and weights $w_n(j)$ satisfy $w_n(j) = w_n(-j) \geqslant 0$,

$$\sum_{|j| \leqslant m_n} w_n(j) = 1$$

and $w_n^2(j) \longrightarrow 0$ as $n \longrightarrow 0$.

Example: $m_n = \sqrt{n}$, $w_n(j) = 1/(2m_n + 1)$ for $|j| \leqslant m_n$.

6.5 Likelihood-based estimation

For parametric models such as ARMA models with Gaussian errors, likelihood-based inference can be carried out using a large sample approximation to the time domain likelihood known as the Whittle likelihood.

Recall that for the periodogram, for any λ, as $n \longrightarrow \infty$

$$\widehat{I}_n(\lambda) \xrightarrow{d} V \sim Exponential\left(1/(2\pi f_X(\lambda))\right)$$

If $\lambda_1, \ldots, \lambda_n$ are chosen to be the Fourier frequencies, then

$$V_1, \ldots, V_n$$

are approximately independent.

6.5 Likelihood-based estimation

The Whittle likelihood is then

$$\mathcal{L}(\theta) = \prod_j \frac{1}{2\pi f_x(\omega_j;\theta)} \exp\left\{ -\frac{\widehat{I}_n(\omega_j)}{2\pi f_x(\omega_j;\theta)} \right\}$$

where $\omega_j = 2\pi j/n$, and

$$j \in \left\{ -\left\lfloor \frac{n-1}{2} \right\rfloor, \ldots, \left\lfloor \frac{n}{2} \right\rfloor \right\}$$

This (asymptotic) likelihood can then be maximized to estimate θ.

6.6 Long memory models

For a stationary ARMA model, recall that the decay in ACF is asymptotically exponentially fast: for h large enough,

$$\rho_X(h) \approx C|\xi_1|^h$$

for some constant C, say, where ξ_1 is reciprocal root of the AR polynomial with modulus closest to 1.

6.6 Long memory models

Let $\{X_t\}$ be a stationary process such that for some $\alpha \in (0,1)$ and constant $c > 0$ such that

$$\lim_{h \longrightarrow \infty} h^{\alpha} \rho_X(h) = c$$

That is, the decay rate is polynomial in h, and therefore relatively slow.

$\{X_t\}$ is a long memory or long range dependent process.

Fractional Integration

Recall the ARIMA(p, d, q) model, for $d = 1, 2, \ldots,$

$$(1 - B)^d \Phi(B) X_t = \Theta(B) Z_t$$

Now suppose that $-0.5 < d < 0.5$ is a non-integer value.

Fractional Integration

We have the binomial expansion for $-0.5 < d < 0.5$

$$(1-B)^d = \sum_{j=0}^{\infty} \psi_j B^j$$

where, for $j = 1, 2, \ldots$

$$\psi_j = \prod_{k=1}^{j} \left(\frac{k-1-d}{k} \right) = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)}$$

where $\Gamma(\cdot)$ is the Gamma function.

Fractional Integration

Let $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, and $W_t = (1-B)^{-d} Z_t$, so that

$$\Phi(B) X_t = \Theta(B) W_t$$

gives an ARMA(p, q) model driven by $\{W_t\}$.

$\{W_t\}$ is fractionally integrated white noise. It is a stationary process with

$$\gamma_W(0) = \frac{\sigma_Z^2 \Gamma(1-2d)}{\{\Gamma(1-d)\}^2}$$

and

$$\rho_W(h) = \frac{\Gamma(h+d)\Gamma(1-d)}{\Gamma(h-d+1)\Gamma(d)} = \prod_{k=1}^{h} \left(\frac{k-1+d}{k-d} \right)$$

Fractional Integration

The ACVF of $\{X_t\}$ can be computed using the usual approach

$$\gamma_X(h) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \psi_j \psi_k \gamma_W(h + j - k)$$

where $\{\psi_j\}$ are the linear process coefficients corresponding to the ARMA process with polynomials $\Phi(B)$ and $\Theta(B)$.

Spectral density for Fractional ARIMA

If $\{X_t\} \sim \text{ARIMA}(p, d, q)$ with $-1/2 < d < 1/2$, then

$$f_X(\lambda) = \frac{\sigma_Z^2}{2\pi} \frac{\left|\Theta\left(e^{-i\lambda}\right)\right|^2}{\left|\Phi\left(e^{-i\lambda}\right)\right|^2} \frac{1}{\left|1 - e^{-i\lambda}\right|^{2d}}$$

If $p = q = 0$,

$$f_X(\lambda) = \frac{\sigma_Z^2}{2\pi} \frac{1}{\left|1 - e^{-i\lambda}\right|^{2d}}$$

and as $\lambda \longrightarrow 0$

$$f_X(\lambda) \approx \frac{\sigma_Z^2}{2\pi\left|\lambda\right|^{2d}}$$

that is, the SDF is unbounded at zero.

Fractional White Noise

If $\{W_t\} \sim \text{ARIMA}(0, d, 0)$, that is,

$$W_t = (1 - B)^{-d} Z_t$$

then

- MA(∞) representation

$$W_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \qquad \psi_j = \frac{\Gamma(j + d)}{\Gamma(j + 1)\Gamma(d)}$$

- AR(∞) representation

$$Z_t = \sum_{j=0}^{\infty} \pi_j W_{t-j} \qquad \pi_j = \frac{\Gamma(j - d)}{\Gamma(j + 1)\Gamma(-d)}$$

Fractional White Noise

For large j

- MA(∞) representation

$$\psi_j \approx \frac{j^{d-1}}{\Gamma(d)}$$

- AR(∞) representation

$$\pi_j \approx \frac{j^{-d-1}}{\Gamma(-d)}$$

Fractional White Noise

If $\{X_t\} \sim \text{ARIMA}(p, d, q)$, write

$$(1 - B)^d \Phi(B) X_t = \Theta(B) Z_t$$

and suppose $0 < d < 0.5$. Then for large h

$$\gamma_X(h) \approx C|h|^{2d-1}.$$

Chapter 7
Financial Time Series Models

7.1 Volatility Modelling

One form of non-stationarity not yet examined is the relaxation of the constant variance assumption.

We use the general term volatility to capture the concept of time-varying variance of a process. This could be considered to arise in

- the unconditional or marginal model of the observation process
- the conditional model for the process (eg in a state space setting).

7.1 Volatility Modelling

It has long been recognized that financial time series exhibit changes in volatility over time that are serially correlated.

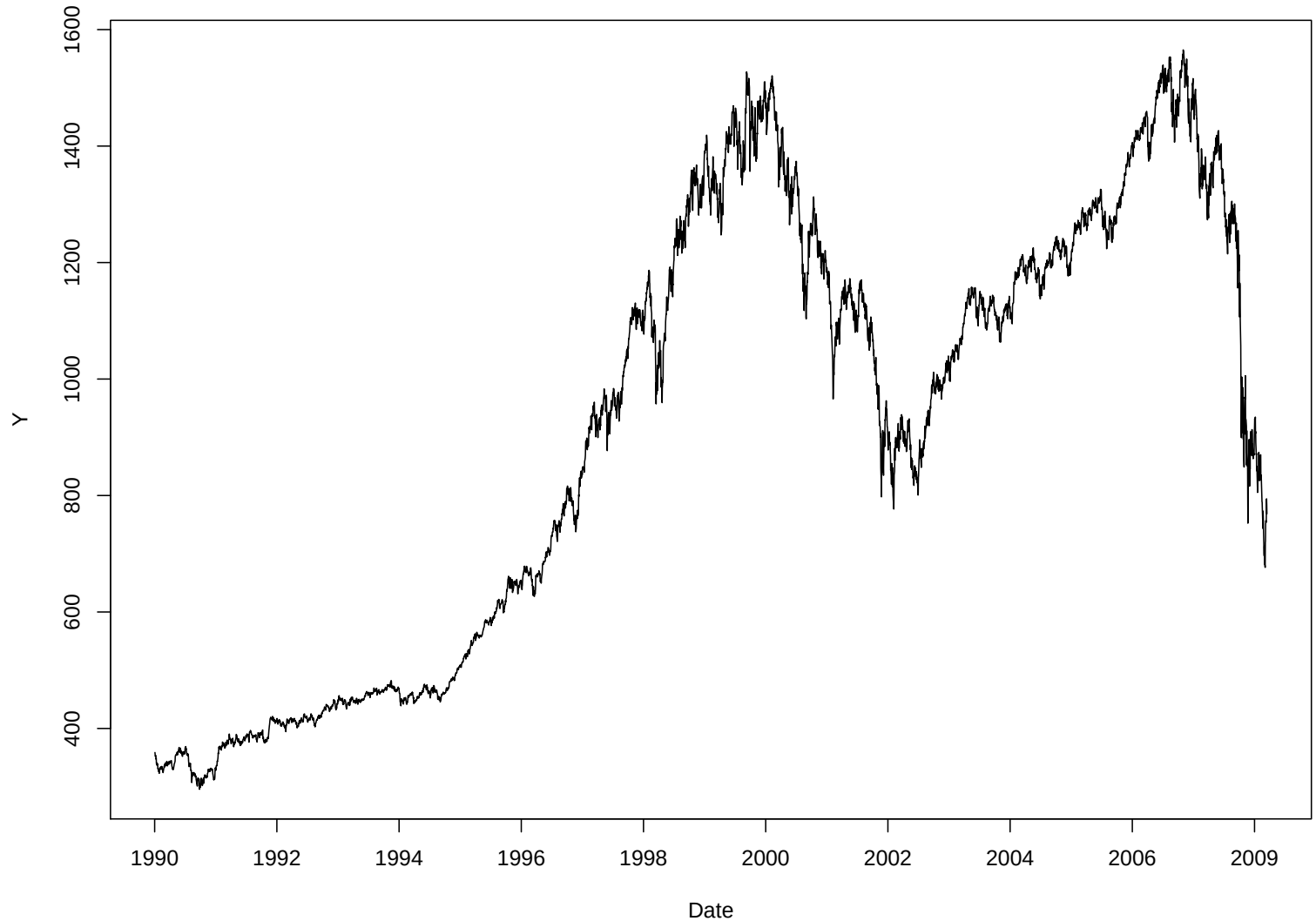
In particular, financial returns demonstrate

volatility clustering

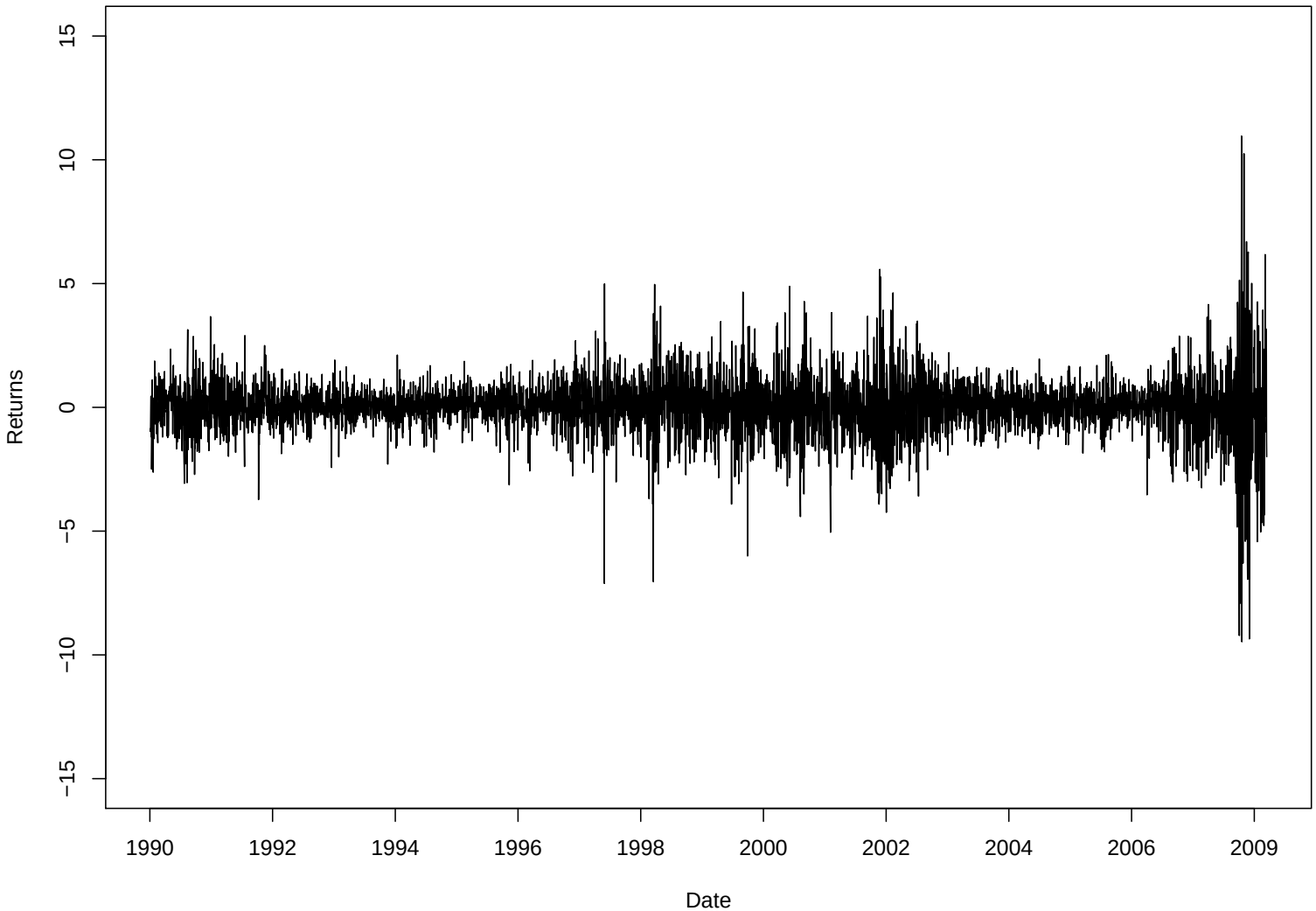
that is, large changes tend to be followed by large changes and vice versa.

Models need to be developed to capture this behaviour.

S & P 500 : 1990-2009



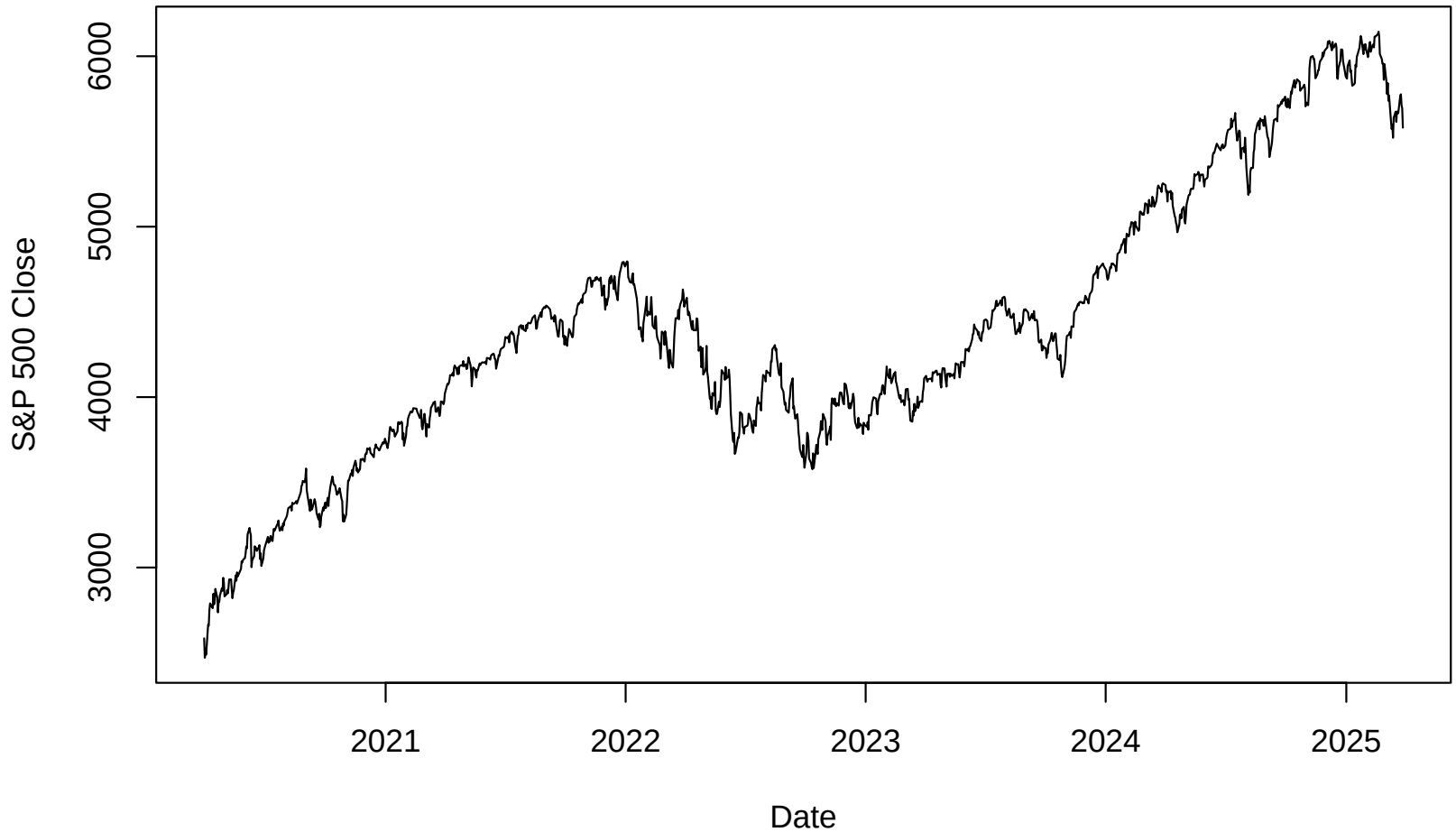
S & P 500 : 1990-2009 log Returns



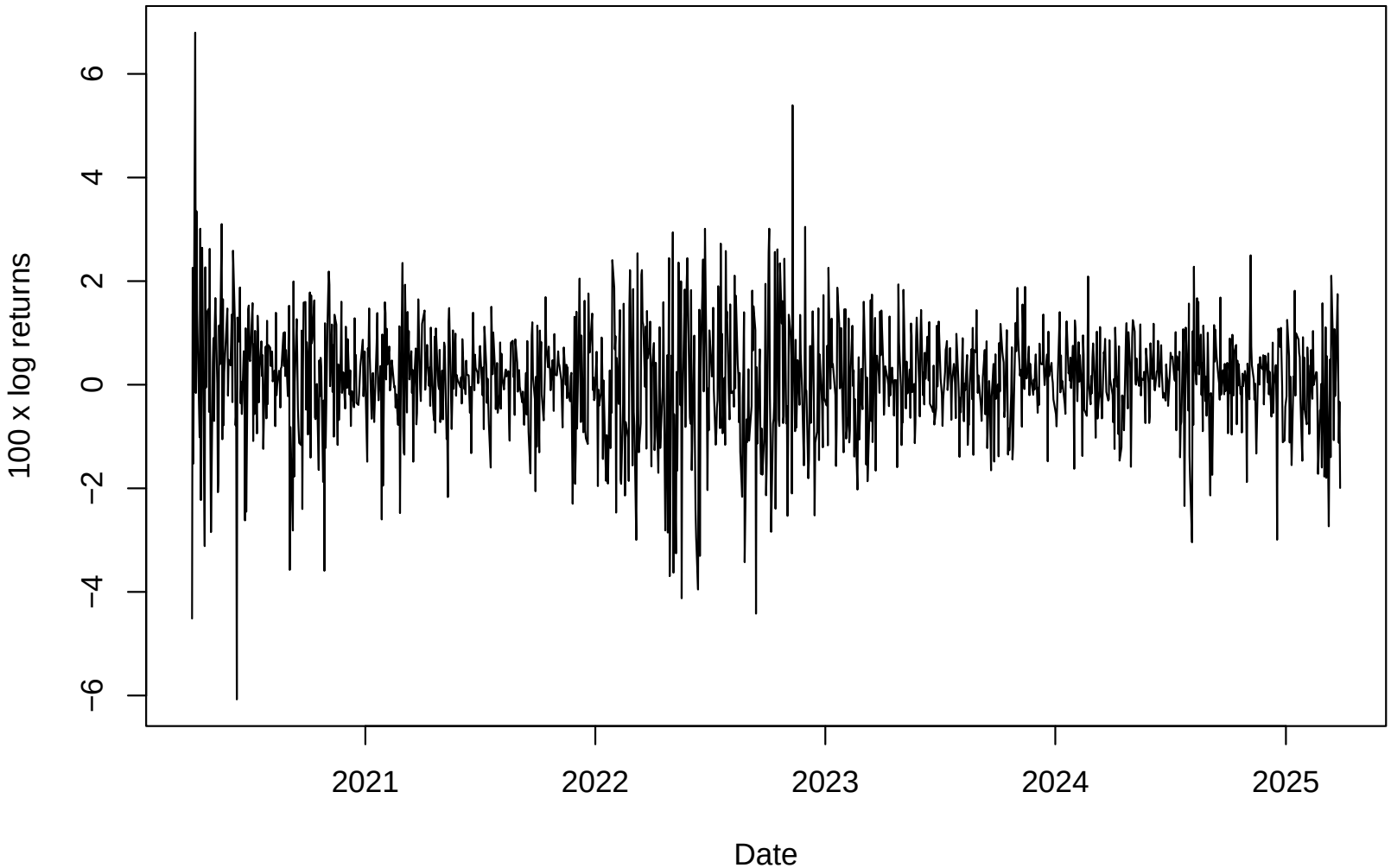
S & P 500 : 2020-2025

```
my.url<-'https://www.math.mcgill.ca/dstephens/545/Data/HistoricalData_1743471926857.csv'
sp25<-read.csv(my.url,header=T)
dvec<-rev(as.character(sp25[,1]))
date.vec<-as.Date(dvec,format="%m/%d/%Y")
Yt<-rev(sp25[,2])
Yt.ret<-diff(log(Yt))
par(mar=c(4,4,2,2))
plot(date.vec,Yt,type='l',xlab='Date',ylab='S&P 500 Close')
```

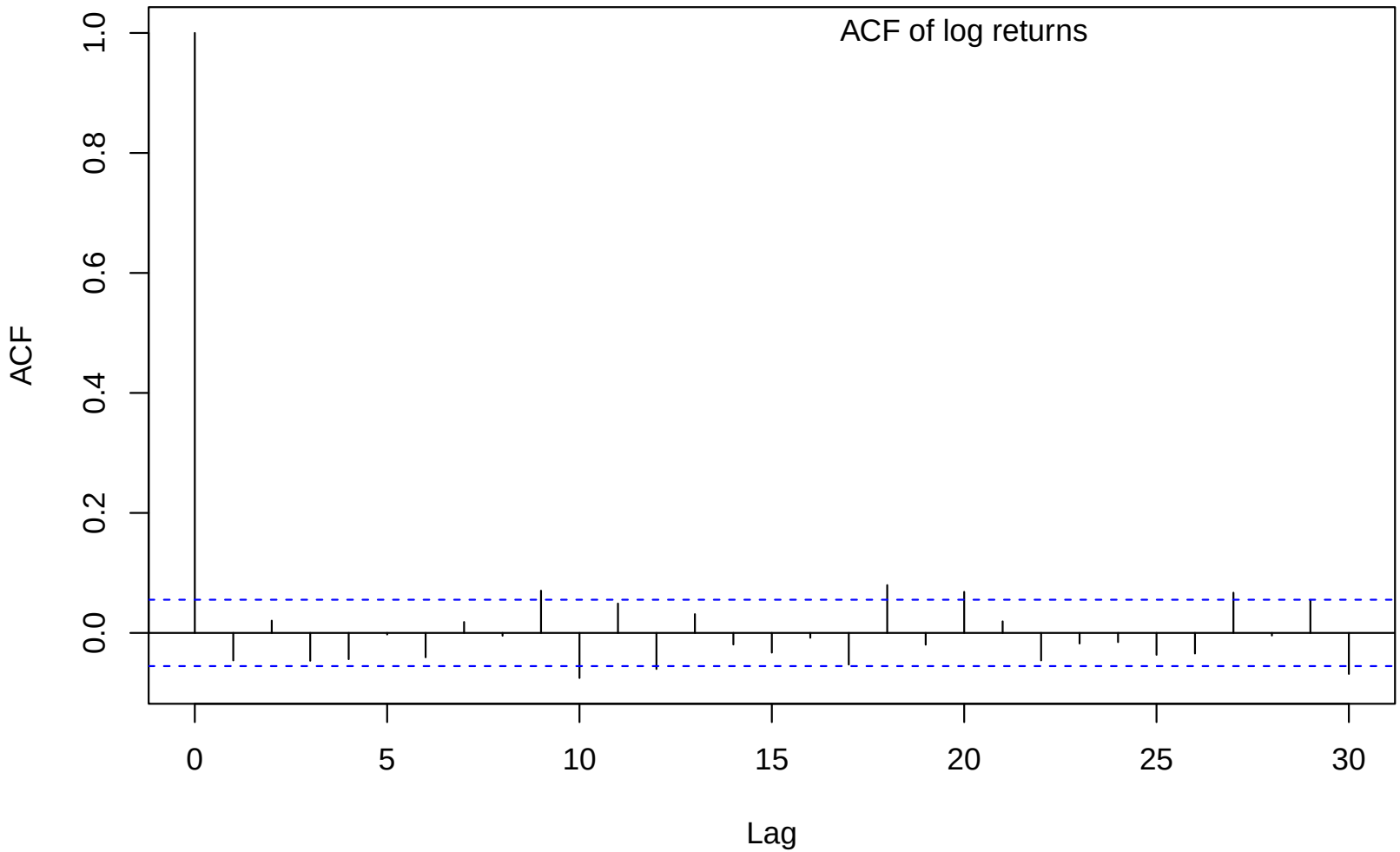
S & P 500 : 2020-2025



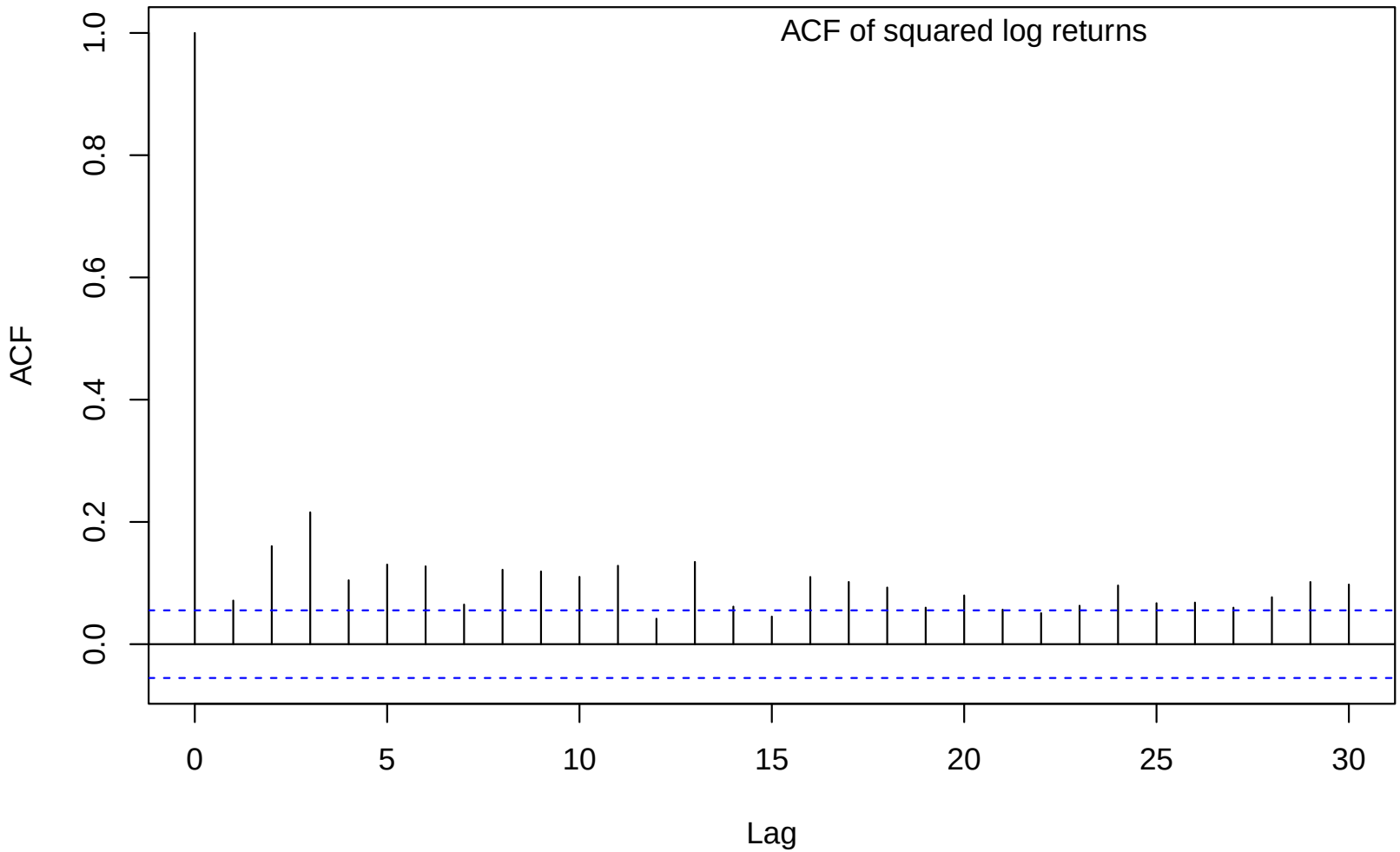
S & P 500 : 2020-2025



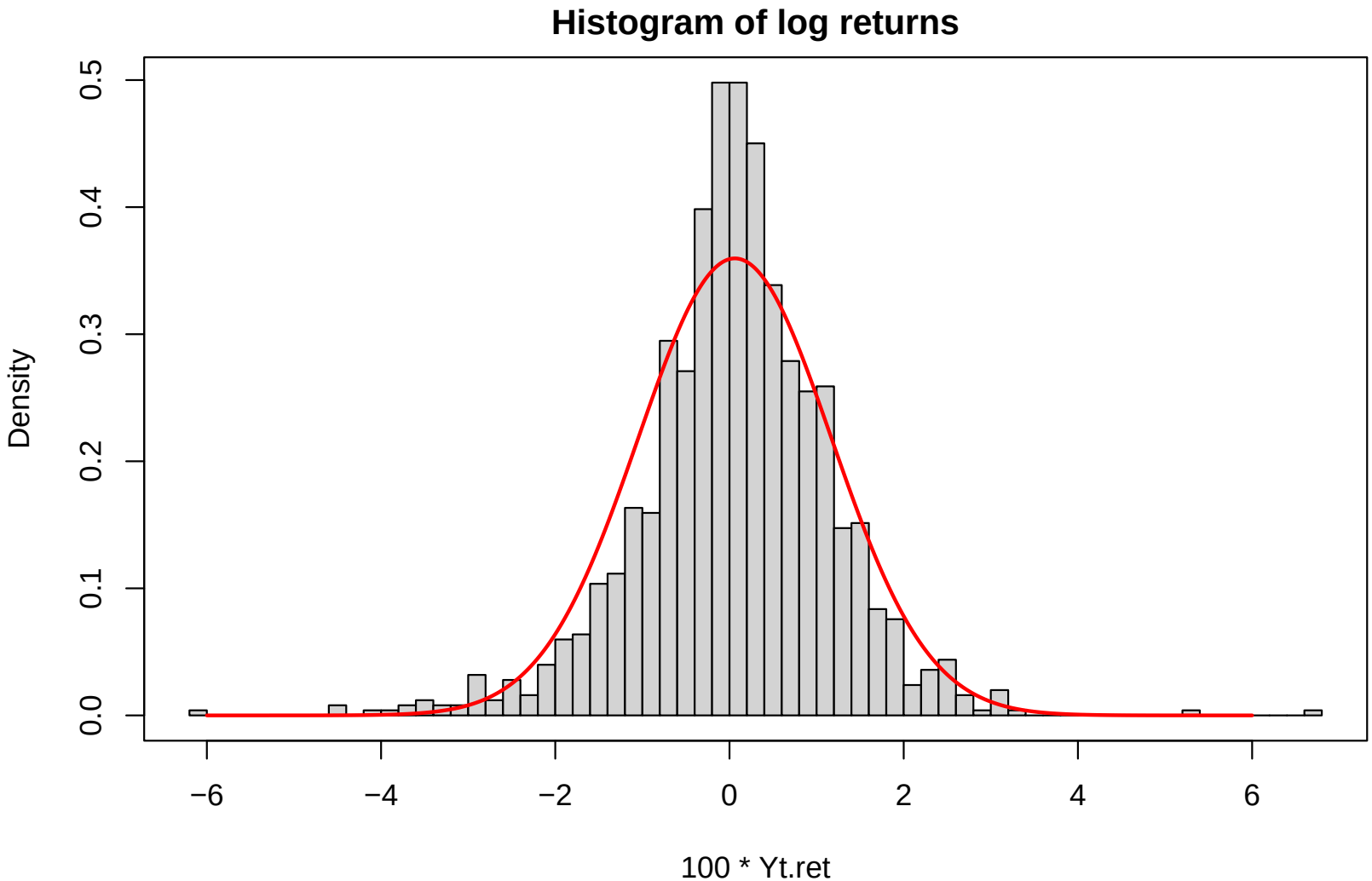
S & P 500 : 2020-2025



S & P 500 : 2020-2025



S & P 500 : 2020-2025



7.2 Types of Volatility Models

Volatility models fall into two classes:

- Observation-driven models allow the variance of the observed series to depend on its lagged values
- Parameter-driven models specify that the variance of the observations is a function of some unobserved or latent process.

These models have the general form of state-space models.

7.2 Types of Volatility Models

In a parameter-driven model

$$\begin{aligned} p\left(y_t \middle| x_t, x_{1:(t-1)}, y_{1:(t-1)}\right) &\equiv p\left(y_t \middle| x_t\right) \\ p\left(x_{t+1} \middle| x_t, x_{1:(t-1)}, y_{1:t}\right) &\equiv p\left(x_{t+1} \middle| x_t\right) \end{aligned}$$

so that

$$p\left(x_{2:n}, y_{1:n} \middle| x_1\right) = \left\{\prod_{t=1}^{n} p\left(y_t \middle| x_t\right)\right\} \left\{\prod_{t=2}^{n} p\left(x_t \middle| x_{t-1}\right)\right\}$$

7.2 Types of Volatility Models

In an observation-driven model

$$\begin{aligned} p(y_t|x_t, x_{1:(t-1)}, y_{1:(t-1)}) &\equiv p(y_t|x_t) \\ p(x_{t+1}|x_t, x_{1:(t-1)}, y_{1:t}) &\equiv p(x_{t+1}|y_{1:t}) \end{aligned}$$

so that

$$p(x_{1:n}, y_{1:n}|x_1) = \prod_{t=1}^{n} p(y_t|x_t)p(x_t|y_{1:(t-1)})$$

where $p(x_1|y_0) \equiv p(x_1)$, say.

7.2 Types of Volatility Models

Example: Model for Count Data

Suppose that for $t = 2, \ldots, n$

$$p(y_t|x_t) \equiv Poisson(\xi x_t)$$

$$p(x_{t+1}|x_t) = \begin{cases} 1 & \text{if } x_{t+1} = a x_t \\ 0 & \text{otherwise} \end{cases}$$

and that where $p(x_1) \equiv Gamma(\alpha, \lambda)$, for some constants ξ, a, α, λ.

7.2 Types of Volatility Models

Example: Model for Count Data

We need to compute the filtering and prediction distributions: Suppose that

$$\begin{aligned} p(x_t|y_{1:t}) &\equiv \text{Gamma}(\alpha_t, \lambda_t) \\ p(x_{t+1}|y_{1:t}) &\equiv \text{Gamma}(\alpha_t, \lambda_t/a) \\ p(y_{t+1}|y_{1:t}) &\equiv \text{NegBinomial}(\alpha_t, 1-\xi/\lambda_t) \end{aligned}$$

where

$$\alpha_t = \alpha + \sum_{j=1}^{t} y_j \qquad \lambda_t = \lambda a^{1-t} + \xi\left(\frac{1-a^{-t}}{1-a^{-1}}\right)$$

7.2 Types of Volatility Models

Example: Model for Count Data

Then

$$\mathbb{E}[Y_{t+1}|Y_{1:t} = y_{1:t}] = \frac{\xi\alpha_t}{\lambda_{t+1} - \xi}$$

and

$$\text{Var}[Y_{t+1}|Y_{1:t} = y_{1:t}] = \frac{\xi\alpha_t\lambda_t}{(\lambda_{t+1} - \xi)^2}$$

7.2 Types of Volatility Models

Example: AR(1) as an observation-driven model

Suppose that

$$\begin{aligned} p(x_{t+1}|y_{1:t}) &= 1 \text{ if } x_{t+1} = y_t, 0 \text{ otherwise} \\ p(y_t|x_t) &\equiv f_Z(y_t - \phi x_t) \end{aligned}$$

for some pdf f_Z. Then

$$Y_t - \phi X_t \stackrel{d}{=} Z_t \qquad \text{or} \qquad Y_t = \phi X_t + Z_t$$

so with $X_t = Y_{t-1}$ we have an AR(1) model.

7.2 Types of Volatility Models

A standard model for financial times series is based on the Stochastic Differential Equation (SDE)

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

where $\{W_t\}$ is a continuous time process with Gaussian increments dW_t.

The equivalent equation for $Y_t = \log S_t$ is

$$dY_t = \mu dt + \sigma dW_t$$

7.2 Types of Volatility Models

Consider the discrete time process where we replace dY_t by $Y_{t+1} - Y_t$, and dt by $\triangle$, with observation times

$$\triangle, 2\triangle, 3\triangle, \ldots$$

to yield data $y_1, y_2, y_3, \ldots$. Note that

$$Y_{t+1} - Y_t = \log S_{t+1} - \log S_t = \log(S_{t+1}/S_t)$$

$$= \mu\triangle + \sigma Z_t$$

where $\{Z_t\} \sim \mathcal{N}(0,1)$, so that

$$Y_{t+1}|Y_t = y_t \sim \mathcal{N}(y_t + \triangle\mu, \sigma^2).$$

7.2 Types of Volatility Models

A better model is based on the SDE

$$dY_t = \mu dt + \sigma(t) dW_t$$

Thus, in the discrete approximation

$$Y_{t+1} - Y_t = \mu \triangle + \sigma_t Z_t$$

where again $\{Z_t\} \sim \mathcal{N}(0, 1)$, so that

$$Y_{t+1} | Y_t = y_t \sim \mathcal{N}(y_t + \triangle\mu, \sigma_t^2)$$

This is referred to as a conditional heteroscedastic model with time-varying variance.

7.2 Types of Volatility Models

The most popular examples of observation-driven models in finance are the

- Autoregressive Conditional Heteroscedasticity (ARCH)
- Generalized ARCH (GARCH).

models.

7.2 Types of Volatility Models

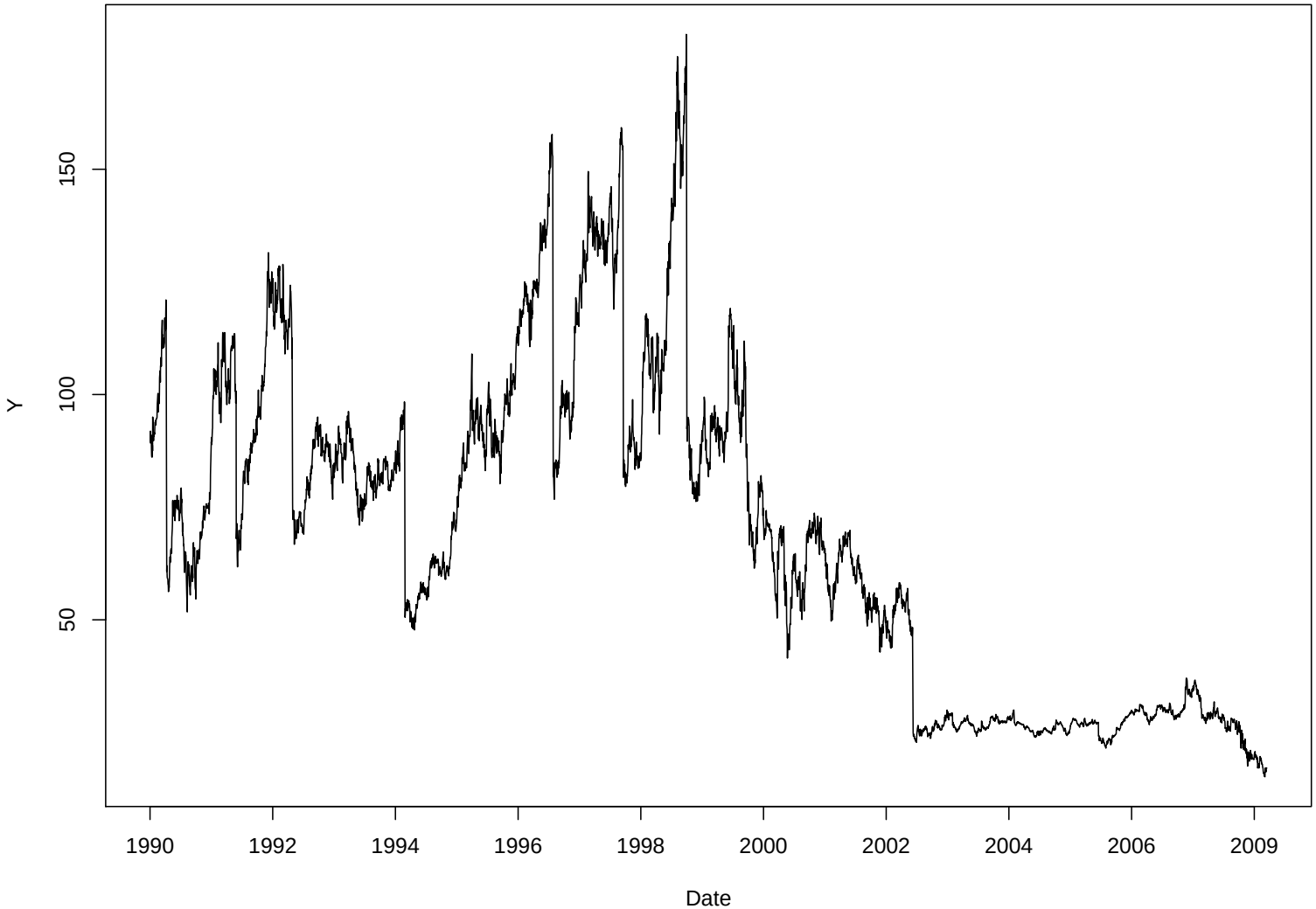
In particular, let y_t be a realization, at time t, of the time series of interest. Typically, y_t is taken to be the compounded return of the underlying asset, so that

$$y_t = 100 \log \left(s_t / s_{t-1} \right)$$

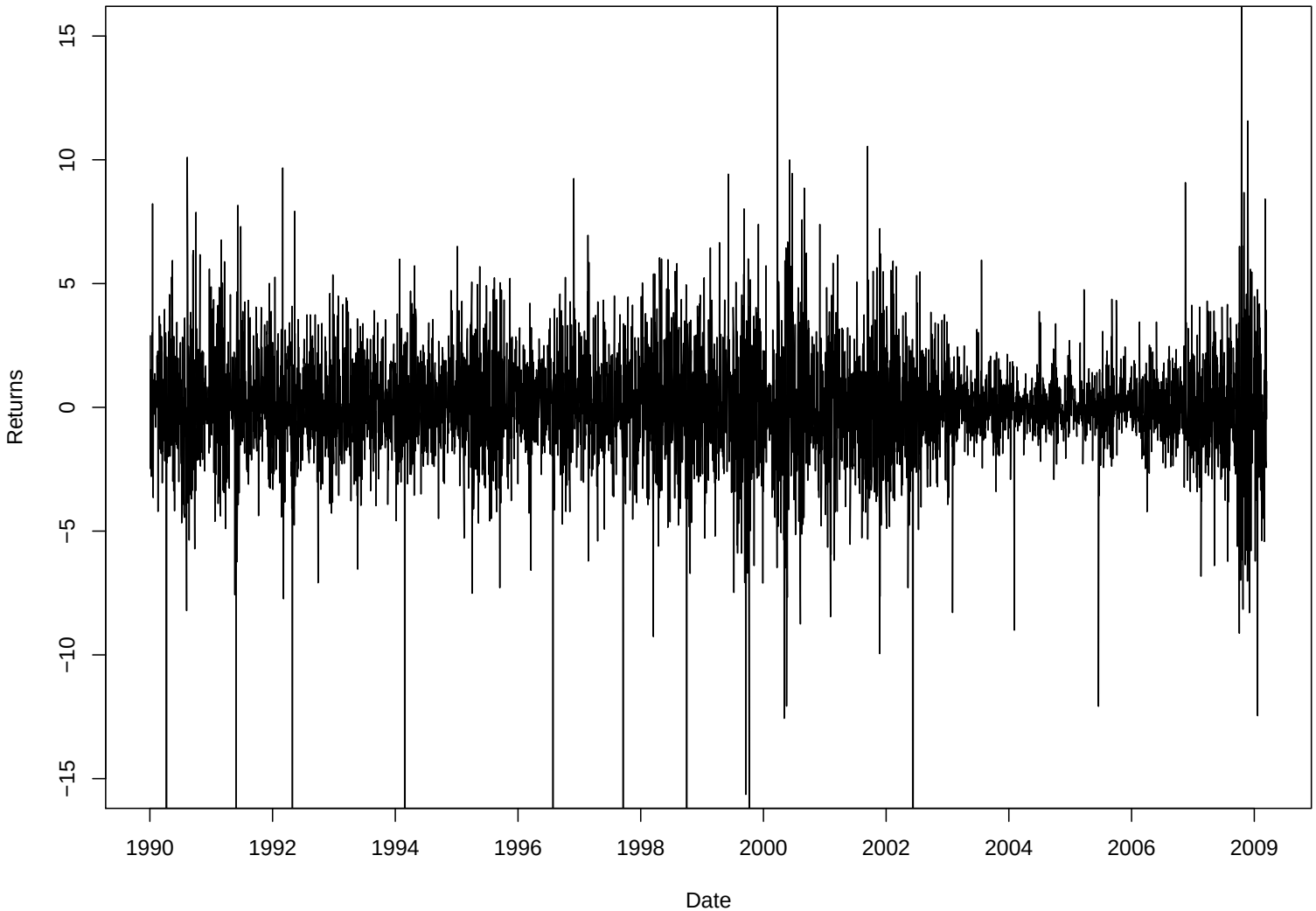
where s_t denotes the observed price of the asset.

ARCH type models specify the distribution of the current observation as a one-step-ahead prediction density.

Microsoft : 1990-2009



Microsoft : 1990-2009 log Returns



7.3 Observation Driven Models

For the observation-driven models,

$$Y_t \mid y_{1:(t-1)} \sim \mathcal{N}\left(0, \sigma_t^2\right)$$

so that

$$Y_t = \sigma_t Z_t,$$

where $\{Z_t\}$ is a sequence of independent $\mathcal{N}(0,1)$ random variables.

7.3 Observation Driven Models

The ARCH(p) model allows the conditional variance σ_t^2 of Y_t to be a linear combination of previous squared observations, so that

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2.$$

Properties of the ARCH(1) model

The simplest ARCH model is the ARCH(1)

$$\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2.$$

Parameters α_0 and α_1 are non-negative, and by iterated expectation, the process is stationary if and only if $\alpha_1 < 1$, with

$$\text{Var}\left[Y_t\right] = \mathbb{E}\left[Y_t^2\right] = \frac{\alpha_0}{1-\alpha_1}.$$

Properties of the ARCH(1) model

All the odd moments of Y_t are zero by symmetry, while the fourth moment exists if and only if $3\alpha_1^2 < 1$ and is

$$\mathbb{E}\left[Y_t^4\right] = \frac{3\alpha_0^2\left(1-\alpha_1^2\right)}{\left(1-\alpha_1\right)^2\left(1-3\alpha_1^2\right)}$$

and the kurtosis (a measure of distributional heavy tailedness)

$$-3 + \frac{\mathbb{E}\left[Y_t^4\right]}{\mathbb{E}\left[Y_t^2\right]^2}$$

is greater than zero and hence the marginal distribution of Y_t is leptokurtic (that is, has fat tails).

Stationarity in ARCH(1) Variance

The ACF of $\{Y_t^2\}$ is given by

$$\rho_{Y^2}(h) = \alpha_1^h,$$

where $h = 0, 1, 2, \ldots$. Note that

- $\{Y_t\}$ is a white noise process,
- $\{Y_t^2\}$ is an autoregressive process that exhibits volatility clustering controlled by α_1.
- $\{Y_t^2\}$ is not necessarily covariance stationary; its variance will be finite only if $3\alpha_1^2 < 1$.

Likelihood Function For ARCH(1) Model

Let y_0 be a fixed constant used to seed the recursion for σ_t. The likelihood for ARCH(1) takes the form

$$p\left(y_{1:n}|y_0,\alpha_0,\alpha_1\right)=\prod_{t=1}^{n}\left(\frac{1}{2\sigma_t^2}\right)^{\frac{1}{2}}\exp\left(-\frac{y_t^2}{2\sigma_t^2}\right),$$

so the log-likelihood is

$$\begin{aligned}\log p\left(y_{1:n}|y_0,\alpha_0,\alpha_1\right) &= \sum_{t=1}^{n}\log p\left(y_t|y_{t-1},\alpha_0,\alpha_1\right)\\ &= const.-\frac{1}{2}\sum_{t=1}^{n}\left[\log\sigma_t^2+\frac{y_t^2}{\sigma_t^2}\right].\end{aligned}$$

GARCH(p, q) model

The GARCH(p, q) process is an extension to the ARCH(p) model which models σ_t^2 as dependent on its lagged values;

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2.$$

The most widely used GARCH model is that of order $(1, 1)$. Sufficient conditions for $\sigma_t^2 \geqslant 0$ are $\alpha_i \geqslant 0$, $i = 0, 1$ and $\beta_1 \geqslant 0$.

The GARCH$(1, 1)$ process has zero mean, and is second order stationary if and only if $\alpha_1 + \beta_1 < 1$, with

$$\text{Var}[Y_t] = \frac{\alpha_0}{(1 - \alpha_1 - \beta_1)}$$

and all the odd moments equal to zero.

GARCH(p, q) model

If in addition,

$$3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$$

the fourth moment exists and is equal to

$$\mathbb{E}\left[Y_t^4\right] = \frac{3\alpha_0^2\left(1 + \alpha_1 + \beta_1\right)}{\left(1 - \alpha_1 - \beta_1\right)\left(1 - 3\alpha_1^2 - 2\alpha_1\beta_1 - \beta_1^2\right)}$$

and again Y_t exhibits leptokurtosis.

A special case of the GARCH$(1, 1)$ model has $\alpha_1 + \beta_1 = 1$, which is called the Integrated GARCH(IGARCH) model.

GARCH models

The moments of the GARCH(1,1) model are:

(i) $\mathbb{E}[Y_t] = \mathbb{E}[Y_t^3] = 0$

(ii) If $0 \leqslant \alpha_1 + \beta_1 < 1$,

$$\mathbb{E}[Y_t^2] = \frac{\alpha_0}{(1 - \alpha_1 - \beta_1)},$$

(iii) If $0 \leqslant \alpha_1 + \beta_1 < 1$ and $3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$

$$\mathbb{E}[Y_t^4] = \frac{3\alpha_0^2 (1 + \alpha_1 + \beta_1)}{(1 - \alpha_1 - \beta_1)(1 - \beta_1^2 - 2\alpha_1\beta_1 - 3\alpha_1^2)},$$

The fourth moment does not exist when the sum of $\alpha_1 + \beta_1$ is close to one, and the value of α_1 is not close to zero.

(iv) If $3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$, the kurtosis is

$$\kappa = \frac{3\left(1+\alpha_1+\beta_1\right)\left(1-\alpha_1-\beta_1\right)}{\left(1-\beta_1^2-2\alpha_1\beta_1-3\alpha_1^2\right)},$$

When $\beta_1 = 0$, this condition is the same as the ARCH(1) model, but when $\beta_1 > 0$, α_1 has to be less than $\sqrt{1/3}$. For example, in the typical case where α_1 is not close to zero and β_1 is near to one, κ does not exist.

GARCH models

(v) The ACF of $\{Y_t^2\}$ is given by

$$\begin{aligned} \rho_{Y^2}(1) &= \frac{\alpha_1(1-\beta_1^2-\alpha_1\beta_1)}{(1-\beta_1^2-2\alpha_1\beta_1)} \\ \rho_{Y^2}(h) &= (\alpha_1+\beta_1)\rho_{Y^2}(h-1), \qquad \text{for } h \geqslant 2. \end{aligned}$$

The ACF declines geometrically at the rate of $\alpha_1 + \beta_1$.

GARCH models

GARCH(p, q) is stationary if the coefficients in the conditional variance equation sum to less than one

$$\alpha_1 + ... + \alpha_p + \beta_1 + ... + \beta_q < 1,$$

in which case the unconditional variance of Y_t,

$$\frac{\alpha_0}{1 - (\alpha_1 + ... + \alpha_p + \beta_1 + ... + \beta_q)}$$

is a finite constant.

GARCH models

For the GARCH$(1,1)$ model: if $\alpha_1 + \beta_1 < 1$, under normality of the residual errors,

$$\text{Var}[Y_t] \quad = \quad const. \frac{\alpha_0}{1 - (\alpha_1 + \beta_1)}$$

and $\text{Cov}[Y_t, Y_s] = 0$, but $\text{Cov}[Y_t^2, Y_s^2] \neq 0$.

ARCH/GARCH model extenstions

There exist many other versions of the ARCH type models,

- Exponential GARCH(EGARCH)
- ARCH-in-Mean (ARCH-M)
- TGARCH
- MGARCH

7.4 Stochastic Volatility Models

The stochastic volatility (SV) model is a form of non-linear state space and parameter-driven model where the variance of the observations to be an unobserved random process.

7.4 Stochastic Volatility Models

The simplest SV model takes the form

$$Y_t = \exp\left(X_t/2\right) Z_t$$
$$X_t = \phi_0 + \phi X_{t-1} + W_t$$

where Y_t is the observation at time t, $Z_t \sim \mathcal{N}(0,1)$, and $W_t \sim \mathcal{N}\left(0, \sigma_W^2\right)$ are IID series.

The latent process $X_t \sim \text{AR}(1)$ is interpreted as the random and uneven flow of new information into the market, and ϕ is the persistence in the volatility.

7.4 Stochastic Volatility Models

The SV model with leverage has

$$Y_t = \exp(X_t/2)\, Z_t$$
$$X_t = \phi_0 + \phi X_{t-1} + W_t$$

but now the two processes $\{Z_t\}$ and $\{W_t\}$ are correlated.

We could have that one of

$$\text{Corr}[Z_t, W_t] = \rho \quad \text{Corr}[Z_{t+1}, W_t] = \rho \quad \text{Corr}[Z_t, W_{t+1}] = \rho$$

holds.

7.4 Stochastic Volatility Models

The parameter ϕ captures the persistence in the volatility, and the restriction

$$|\phi| < 1$$

is typically imposed to ensure that the series X_t of the log-volatilities is stationary.

However, X_t can also be allowed to follow a random walk by setting $\phi = 1$.

Properties of the Stochastic Volatility Model

Suppose $\{Z_t\}$ and $\{W_t\}$ are independent. Under stationarity, the mean and variance of $\{X_t\}$ given respectively by

$$\mu_X = \mathbb{E}[X_t] = \frac{\phi_0}{1-\phi} \qquad \sigma_X^2 = \text{Var}[X_t] = \frac{\sigma_W^2}{1-\phi^2}.$$

Since

$$Y_t = Z_t \exp(X_t/2)$$

and $\{Z_t\}$ is always stationary, $\{Y_t\}$ will also be stationary if and only if $\{X_t\}$ is stationary. If the distribution of Z_t is symmetric, then all the odd moments of Y_t are zero.

Properties of the Stochastic Volatility Model

By iterated expectation

$$\mathbb{E}\left[Y_t\right] = \mathbb{E}\left[\mathbb{E}\left[Y_t \mid Y_{1:(t-1)}\right]\right] = 0$$

so that Y_t is zero mean and the ACF of Y_t is

$$\gamma_Y\left(h\right) = \mathbb{E}\left[Y_t Y_{t-h}\right] = \mathbb{E}\left[\exp\left(\frac{X_t}{2} + \frac{X_{t-h}}{2}\right)\right]\mathbb{E}\left[Z_t Z_{t-h}\right] = 0$$

for $h \geqslant 1$, as $\left\{Z_t\right\}$ is an uncorrelated process.

Properties of the Stochastic Volatility Model

By assumption, $\exp(X_t)$ has a log-normal distributed, and

$$\mathbb{E}\left[\exp(X_t)^j\right] = \exp\left\{j\mu_X + \frac{1}{2}j^2\sigma_X^2\right\},$$

so that, if $r = 2s$ is even and $\{X_t\}$ is stationary, all the even moments of Y_t exist and are given by the formula

$$\mathbb{E}\left[Y_t^r\right] = \mathbb{E}\left[(\exp(X_t))^s\right]\mathbb{E}\left[Z_t^r\right] = \exp\left\{s\mu_X + \frac{s^2}{2}\sigma_X^2\right\}\frac{(2s)!}{2^s\ s!}.$$

Then

$$\mathrm{Var}\left[Y_t\right] = \mathbb{E}\left[Y_t^2\right] = \exp\left\{\mu_X + \frac{1}{2}\sigma_X^2\right\}$$

Hence if $\{X_t\}$ is stationary, $\{Y_t\}$ is an uncorrelated process.

Properties of the Stochastic Volatility Model

Also

$$\mathbb{E}\left[Y_t^4\right] = 3\exp\left\{2\mu_X + 2\sigma_X^2\right\}$$

and so the kurtosis for Y_t is

$$-3 + \mathbb{E}\left[Y_t^4\right]/\mathbb{E}\left[Y_t^2\right]^2 = 3\left(\exp\left(\sigma_X^2\right) - 1\right),$$

which is greater than 0 if σ_X^2 is positive. Thus, Y_t has a leptokurtic, symmetric distribution.

Linearization

If we square Y_t and take logarithms, then

$$\log Y_t^2 = X_t + \log Z_t^2.$$

If Z_t has a standard normal distribution, then $\log Z_t^2$ has a log-chi-square distribution with mean $\psi(1) - \log 2 \simeq -1.2704$ and variance $\pi^2/2 \doteq 4.9348$, where $\psi(\cdot)$ is the digamma function

$$\psi(z) = \frac{d}{dz} \log \Gamma(z) = \frac{\dot{\Gamma}(z)}{\Gamma(z)}.$$

Linearization

Thus, if $V_t = \log Z_t^2 + 1.2704$, then $\{V_t\}$ is IID with mean zero and variance $\pi^2/2$, we may rewrite the model as

$$\log Y_t^2 = -1.2704 + X_t + V_t.$$

Linearization

Hence $\{\log Y_t^2\}$ is a linear process which is the sum of

- the AR(1) process $\{X_t\}$
- a white noise process.

and $\log Y_t^2$ behaves (approximately) as an ARMA$(1,1)$ process, with ACF given by

$$\rho_{\log Y_t^2}(h) = \frac{\phi^h}{1 + \left(\pi^2/2\sigma_X^2\right)}, \quad h = 1, 2, \ldots.$$

7.5 Inference for the SV Model

No analytic expression exists for the densities $p\left(y_t \mid y_{1:(t-1)}\right)$, and this makes the likelihood function hard to evaluate; the distribution of Y_t conditional on past information $y_{1:(t-1)}$ does not possess an analytic expression.

The likelihood can be computed by integrating the latent log-volatilities out of the joint probability distribution.

7.5 Inference for the SV Model

The likelihood is

$$p(y_{1:n};\theta) = \int p(y_{1:n}, x_{1:n};\theta)\,dx_{1:n}$$

$$= \int p(y_{1:n} \mid x_{1:n};\theta)\,p(x_{1:n};\theta)\,dx_{1:n}.$$

This last integral requires the use of numerical procedures and this makes the estimation θ quite involved.

Generalized Method-of-Moments (GMM)

The generalized method of moments (GMM) requires the construction of a vector $\mathbf{g}$, whose elements will be the differences between the unconditional expectations and the sample moments.

For the SV model there are three parameters we need to estimate, namely $\boldsymbol{\theta} = (\phi_0, \phi, \sigma_W^2)$, and a large number of moments to use.

Generalized Method-of-Moments (GMM)

For example, we could define $\mathbf{g}$ via

$$\begin{aligned}
&\frac{1}{n}\sum y_t^2 - \mathbb{E}\left[Y_t^2\right] \\
&\frac{1}{n}\sum y_t^4 - \mathbb{E}\left[Y_t^4\right] \\
&\frac{1}{n}\sum y_t^2 y_{t-1}^2 - \mathbb{E}\left[Y_t^2 Y_{t-1}^2\right] \\
&\vdots \\
&\frac{1}{n}\sum y_t^2 y_{t-h}^2 - \mathbb{E}\left[Y_t^2 Y_{t-h}^2\right]
\end{aligned}$$

where the theoretical values of $\mathbb{E}[Y_t^2 Y_{t-h}^2]$, for $h \geqslant 1$, can be found analytically.

Generalized Method-of-Moments (GMM)

The objective function to be minimized is then

$$Q = \mathbf{g}^\top \mathbf{W} \mathbf{g}$$

where $\mathbf{W}$ is a positive definite, symmetric weighting matrix.

GMM estimators are consistent if the observed process $\{Y_t\}$ is stationary. However,

- when $|\phi|$ is near 1, the GMM estimator can work poorly.
- the method does not deliver filtered or smoothed estimates of X_t.

7.6 Quasi-Maximum Likelihood estimation

The Quasi-Maximum Likelihood (QML) method is based on the linearization approach described earlier.

Let $Y_t^* = \log Y_t^2$. Then the SV model can be written as

$$Y_t^* = -1.2704 + X_t + V_t$$
$$X_t = \phi_0 + \phi X_{t-1} + W_t$$

where

$$V_t = \log Z_t^2 - \mathbb{E}\left[\log Z_t^2\right]$$

with $\sigma_V^2 = \mathrm{Var}\left[V_t\right] = \pi^2/2$.

7.6 Quasi-Maximum Likelihood estimation

This is a linear but non-Gaussian state-space model. The QML approach treats the observation errors as if

$$V_t \sim \mathcal{N}\left(0, \pi^2/2\right)$$

and applies the standard Kalman filter

Given transformed data $\{y_1^*, \ldots, y_n^*\}$, the Kalman recursions can be used to construct the approximate Gaussian likelihood.

7.7 Bayesian Inference

In a Bayesian procedure, we update our prior belief about $\boldsymbol{\theta}$ and obtain a posterior distribution of the parameters, which is denoted $\pi_n\left(\boldsymbol{\theta}\left|y_{1:n}\right.\right)$ and is given by

$$\pi_n\left(y_{1:n}|\boldsymbol{\theta}\right)=\frac{p\left(y_{1:n}|\boldsymbol{\theta}\right)\pi_0\left(\boldsymbol{\theta}\right)}{\int p\left(y_{1:n}|\mathbf{t}\right)\pi_0\left(\mathbf{t}\right)d\mathbf{t}}\propto p\left(y_{1:n}|\boldsymbol{\theta}\right)\pi_0\left(\boldsymbol{\theta}\right),$$

In the SV case, we can explore the posterior distribution using Markov chain Monte Carlo (MCMC).

7.7 Bayesian Inference

Let

- $\boldsymbol{\theta} = \left(\phi_0, \phi, \sigma_W^2\right)^\top$ denote the vector of parameters,
- $X_{1:n} = (X_1, \ldots, X_n)^\top$ denote the vector of log-volatilities
- $y_{1:n} = (y_1, \ldots, y_n)^\top$ denote the vector of observations,

then the model is specified by the sequence of three conditional distributions.

7.7 Bayesian Inference

Key elements are

- the distribution of the observations conditional on the log-volatilities, $p\left(y_{1:n}|X_{1:n}\right)$,
- the distribution of the log-volatilities conditional on the hyperparameters, $p\left(X_{1:n}|\boldsymbol{\theta}\right)$
- the prior distribution of the parameters, $\pi_0(\boldsymbol{\theta})$.

The joint posterior distribution of the log-volatilities and hyperparameters is

$$\pi_n\left(X_{1:n}, \boldsymbol{\theta}|y_{1:n}\right) \propto p\left(y_{1:n}|X_{1:n}\right) p\left(X_{1:n}|\boldsymbol{\theta}\right) \pi_0(\boldsymbol{\theta}).$$

7.7 Bayesian Inference

The estimation of the SV model via MCMC considers the hierarchical structure of conditional distributions.

MCMC samples iteratively from the conditional posterior distributions: we initialize the values

$$x_{1:n}^{(1)}, \boldsymbol{\theta}^{(1)}$$

and then for iteration $s = 2, 3, \ldots,$ sample

- $\boldsymbol{\theta}^{(s)}$ from $\pi_n(\boldsymbol{\theta} \,|X_{1:n}, y_{1:n})$, assuming $X_{1:n} = x_{1:n}^{(s-1)}$
- $x_{1:n}^{(s)}$ from $\pi_n(X_{1:n}|\boldsymbol{\theta}, y_{1:n})$, assuming $\boldsymbol{\theta} = \boldsymbol{\theta}^{(s)}$

and store the samples $\{\boldsymbol{\theta}^{(s)}, x_{1:n}^{(s)}\}$ over many iterations.

7.7 Bayesian Inference

Particle filtering methods may also be used to perform filtering based inference;

- see previous `knitr` sheet.

It is also possible to integrate particle methods into MCMC:

- particle methods are used to estimate the likelihood
- the estimated likelihood used in the MCMC calculations
- can be effective if the number of particles is large.