Introduction to Time Series

Jean-Paul Renne

2023-01-13

Contents

Introduction to Time Series

Time series constitute a prevalent data type in several disciplines, notably macroeconomics and finance. The modeling of time series is crucial for many purposes, including forecasting, understanding macroeconomic mechanisms, and risk assessment. This course proposes an introduction to time series analysis. It has been developed by Jean-Paul Renne.

Codes associated with this course are part of the `AEC` package, which is available on GitHub. To load a package from GitHub, you need to use function `install_github` from the `devtools` package:

```
install.packages("devtools") # in case you do not have that one.
library(devtools)
install_github("jrenne/AEC")
library(AEC)
```

Useful (R) links:

- Download R:
  - R software: https://cran.r-project.org (the basic R software)
  - RStudio: https://www.rstudio.com (a convenient R editor)
- Tutorials:
  - Rstudio: https://dss.princeton.edu/training/RStudio101.pdf (by Oscar Torres-Reyna)
  - R: https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf (by Emmanuel Paradis)
  - My own tutorial: https://jrenne.shinyapps.io/Rtuto_publiShiny/

Chapter 1

To start with

1.1 Shocks and lag operator

A time series is an infinite sequence of random variables indexed by time: $\{y_t\}_{t=-\infty}^{+\infty} = \{..., y_{-2}, y_{-1}, y_0, y_1, ..., y_t, ...\}$, $y_i \in \mathbb{R}^k$. In practice, we only observe samples, typically: $\{y_1, ..., y_T\}$.

Standard time series models are built using **shocks** that we will often denote by ε_t. Typically, $\mathbb{E}(\varepsilon_t) = 0$. In many models, the shocks are supposed to be i.i.d., but there exist other (less restrictive) notions of shocks. In particular, the definition of many processes is based on white noises:

Definition 1.1 (White noise)**.** The process $\{\varepsilon_t\}_{t\in]-\infty,+\infty[}$ is a white noise if, for all t:

a. $\mathbb{E}(\varepsilon_t) = 0$,
b. $\mathbb{E}(\varepsilon_t^2) = \sigma^2 < \infty$,
c. for all $s \neq t$, $\mathbb{E}(\varepsilon_t \varepsilon_s) = 0$.

Another type of shocks that are commonly used are Martingale Difference Sequences:

Definition 1.2 (Martingale Difference Sequence)**.** The process $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ is a martingale difference sequence (MDS) if $\mathbb{E}(|\varepsilon_t|) < \infty$ and if, for all t,

$$\underbrace{\mathbb{E}_{t-1}(\varepsilon_t)}_{\text{conditional on the past}} = 0.$$

By definition, if y_t is a martingale, then $y_t - y_{t-1}$ is a MDS.

Example 1.1 (ARCH process)**.** The Autoregressive conditional heteroskedasticity process—studied in Section 7—is an example of shock that satisfies the MDS definition but that is not i.i.d.:

$$\varepsilon_t = \sigma_t \times z_t,$$

where $z_t \sim i.i.d.\, \mathcal{N}(0,1)$ and $\sigma_t^2 = w + \alpha\varepsilon_{t-1}^2$.

Example 1.2. A white noise process is not necessarily a MDS. This is for instance the case for following process:

$$\varepsilon_t = z_t + z_{t-1}z_{t-2},$$

where $z_t \sim i.i.d.\mathcal{N}(0,1)$.

Let us now introduce the lag operator. The lag operator, denoted by L, is defined on the time series space and is defined by:

$$L : \{y_t\}_{t=-\infty}^{+\infty} \to \{w_t\}_{t=-\infty}^{+\infty} \quad \text{with} \quad w_t = y_{t-1}. \tag{1.1}$$

It is easily seen that we have $L^2 y_t = L(Ly_t) = y_{t-2}$ and, more generally, $L^k y_t = y_{t-k}$.

Consider a process that follows y_t defined by $y_t = \mu + \phi y_{t-1} + \varepsilon_t$ (this is an AR(1) process, as we will see in Section 2), where the ε_t's are i.i.d. $\mathcal{N}(0, \sigma^2)$. Using the lag operator, the dynamics of y_t can be expressed as follows:

$$(1 - \phi L)y_t = \mu + \varepsilon_t.$$

1.2 Conditional and unconditional moments

If it exists, the **unconditional (or marginal) mean** of the random variable y_t is given by:

$$\mu_t := \mathbb{E}(y_t) = \int_{-\infty}^{\infty} y_t f_{Y_t}(y_t) dy_t,$$

where f_{Y_t} is the unconditional, or marginal, density (p.d.f.) of y_t. Note that, in the general case, Y_t and Y_{t-1}, may have different densities; that is, in general, $f_{Y_t} \neq f_{Y_{t-1}}$.

Similarly, if it exists, the **unconditional (or marginal) variance** of the random variable y_t is:

$$\mathbb{V}ar(y_t) = \int_{-\infty}^{\infty} (y_t - \mathbb{E}(y_t))^2 f_{Y_t}(y_t) dy_t.$$

Definition 1.3 (Autocovariance)**.** The j^{th} autocovariance of y_t is given by:

$$\begin{aligned} \gamma_{j,t} &:= \mathbb{E}([y_t - \mathbb{E}(y_t)][y_{t-j} - \mathbb{E}(y_{t-j})]) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} [y_t - \mathbb{E}(y_t)][y_{t-j} - \mathbb{E}(y_{t-j})] \times \\ &\quad f_{Y_t, Y_{t-1}, ..., Y_{t-j}}(y_t, y_{t-1}, ..., y_{t-j}) dy_t dy_{t-1} ... dy_{t-j}, \end{aligned}$$

where $f_{Y_t, Y_{t-1}, ..., Y_{t-j}}(y_t, y_{t-1}, ..., y_{t-j})$ is the joint distribution of $y_t, y_{t-1}, ..., y_{t-j}$.

In particular, note that $\gamma_{0,t} = \mathbb{V}ar(y_t)$.

Definition 1.4 (Covariance stationarity)**.** The process y_t is covariance stationary —or weakly stationary— if, for all t and j,

$$\mathbb{E}(y_t) = \mu \quad \text{and} \quad \mathbb{E}\{(y_t - \mu)(y_{t-j} - \mu)\} = \gamma_j.$$

Figure **??** displays the simulation of a process that is not covariance stationary. This process follows $y_t = 0.1t + \varepsilon_t$, where $\varepsilon_t \sim \ i.i.d. \ \mathcal{N}(0,1)$. Indeed, for such a process, we have: $\mathbb{E}(y_t) = 0.1t$, which depends on t.

Definition 1.5 (Strict stationarity)**.** The process y_t is strictly stationary if, for all t and all sets of integers $J = \{j_1, ..., j_n\}$, the distribution of $(y_t, y_{t+j_1}, ..., y_{t+j_n})$ depends on J but not on t.

The following process is covariance stationary but not strictly stationary:

$$y_t = \mathbb{I}_{\{t<1000\}} \varepsilon_{1,t} + \mathbb{I}_{\{t \geq 1000\}} \varepsilon_{2,t},$$

where $\varepsilon_{1,t} \sim \mathcal{N}(0,1)$ and $\varepsilon_{2,t} \sim \sqrt{\frac{\nu-2}{\nu}} t(\nu)$ and $\nu = 4$.

Proposition 1.1. *If y_t is covariance stationary, then $\gamma_j = \gamma_{-j}$.*

Proof. Since y_t is covariance stationary, the covariance between y_t and y_{t-j} (i.e γ_j) is the same as that between y_{t+j} and y_{t+j-j} (i.e. γ_{-j}). □

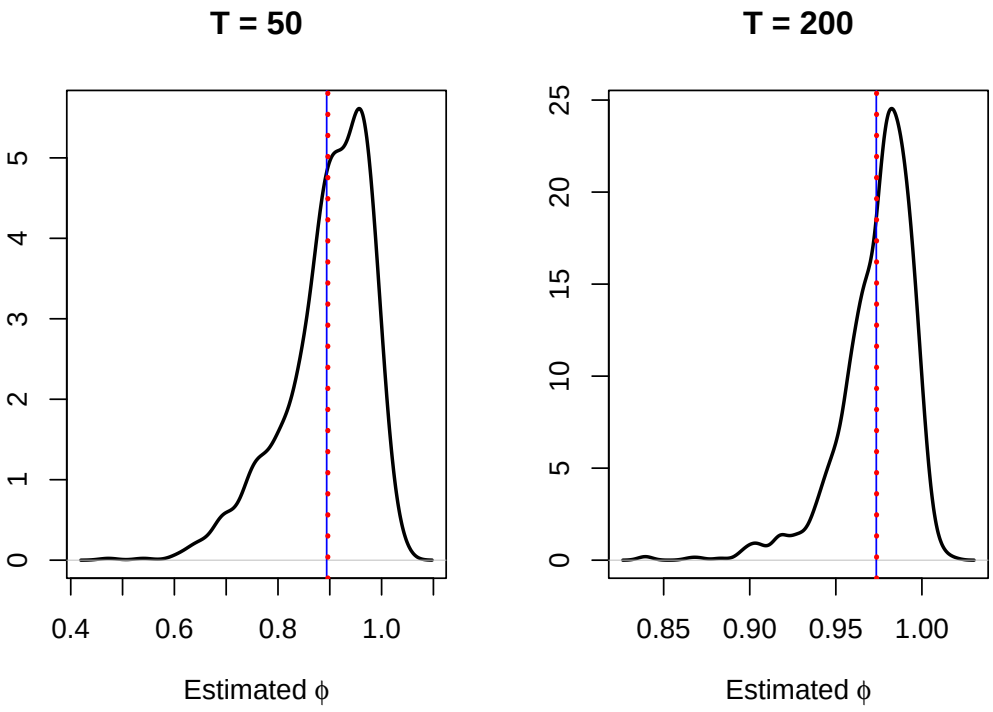


Figure 1.1: Example of a process that is not covariance stationary ($y_t = 0.1t + \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0, 1)$).

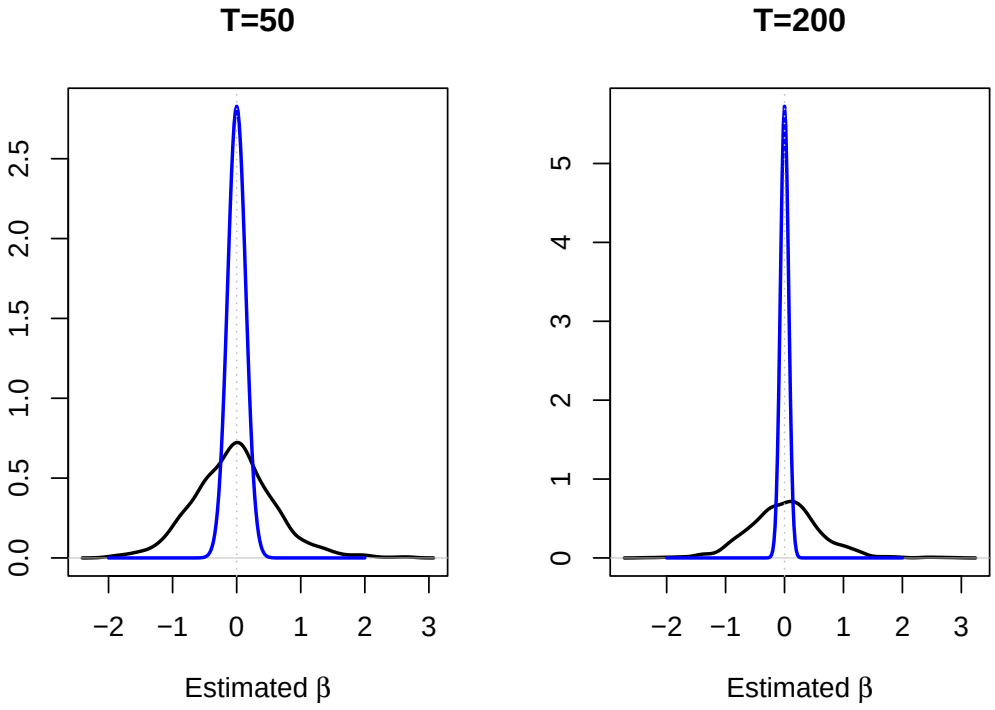


Figure 1.2: Example of a process that is covariance stationary but not strictly stationary. The red lines delineate the 99% confidence interval of the standard normal distribution (± 2.58).

Definition 1.6 (Auto-correlation)**.** The j^{th} auto-correlation of a covariance-stationary process is:

$$\rho_j = \frac{\gamma_j}{\gamma_0}.$$

Consider a long historical time series of the Swiss GDP growth, taken from the Jordà et al. (2017) dataset.[1]

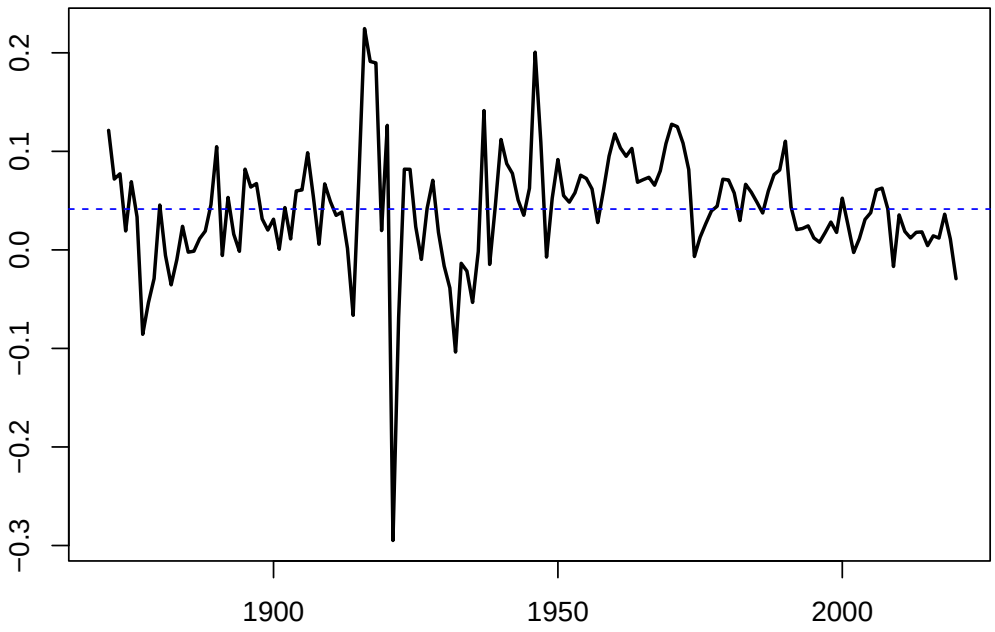


Figure 1.3: Annual growth rate of Swiss GDP, based on the Jorda-Schularick-Taylor Macrohistory Database.

Definition 1.7 (Mean ergodicity)**.** The covariance-stationary process y_t is ergodic for the mean if:

$$\text{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=1}^{T} y_t = \mathbb{E}(y_t).$$

Definition 1.8 (Second-moment ergodicity)**.** The covariance-stationary process y_t is ergodic for second moments if, for all j:

$$\text{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=1}^{T}(y_t-\mu)(y_{t-j}-\mu) = \gamma_j.$$

[1] Version 6 of the dataset, available on this website.

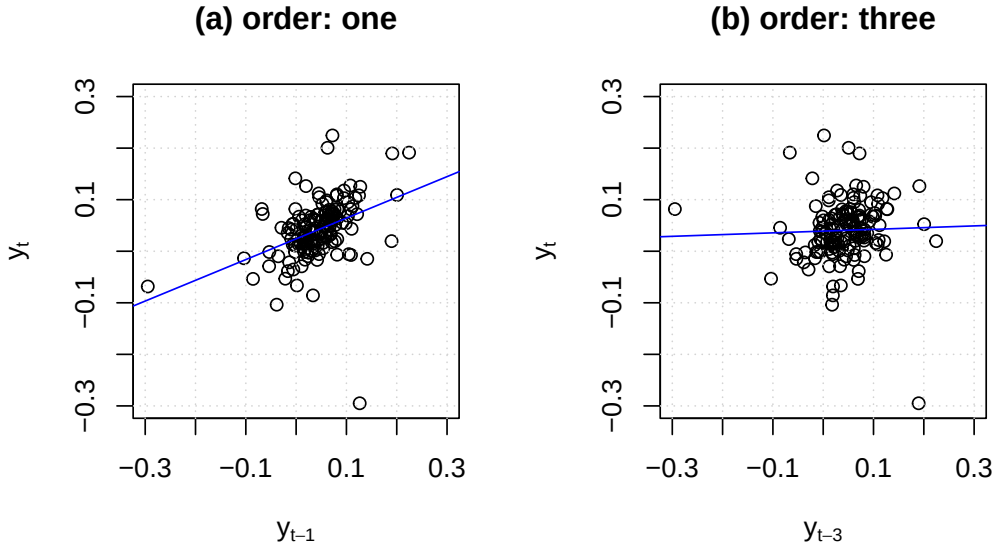


Figure 1.4: For order j, the slope of the blue line is, approximately, $\hat{\gamma}_j/\widehat{\mathbb{V}ar}(y_t)$, where hats indicate sample moments.

It should be noted that ergodicity and stationarity are different properties. Typically if the process $\{x_t\}$ is such that, $\forall t$, $x_t \equiv y$, where $y \sim \mathcal{N}(0,1)$ (say), then $\{x_t\}$ is stationary but not ergodic.

Theorem 1.1 (Central Limit Theorem for covariance-stationary processes)**.** *If process y_t is covariance stationary and if the series of autocovariances is absolutely summable ($\sum_{j=-\infty}^{+\infty} |\gamma_j| < \infty$), then:*

$$\bar{y}_T \overset{m.s.}{\to} \mu = \mathbb{E}(y_t) \tag{1.2}$$

$$lim_{T\to+\infty} T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] = \sum_{j=-\infty}^{+\infty} \gamma_j \tag{1.3}$$

$$\sqrt{T}(\bar{y}_T - \mu) \overset{d}{\to} \mathcal{N}\left(0, \sum_{j=-\infty}^{+\infty} \gamma_j\right). \tag{1.4}$$

[Mean square (m.s.) and distribution (d.) convergences: see Definitions 8.19 and 8.17.]

Proof. By Proposition 8.8, Eq. (1.3) implies Eq. (1.2). For Eq. (1.3), see Appendix 8.5. For Eq. (1.4), see Anderson (1971), p. 429. □

Definition 1.9 (Long-run variance)**.** Under the assumptions of Theorem 1.1, the limit appearing in Eq. (1.3) exists and is called **long-run variance**. It is denoted by S, i.e.:

$$S = \Sigma_{j=-\infty}^{+\infty} \gamma_j = \lim_{T \to +\infty} T \mathbb{E}[(\bar{y}_T - \mu)^2].$$

If y_t is ergodic for second moments (see Def. 1.8), a natural estimator of S is:

$$\hat{\gamma}_0 + 2 \sum_{\nu=1}^{q} \hat{\gamma}_\nu, \tag{1.5}$$

where $\hat{\gamma}_\nu = \frac{1}{T} \sum_{\nu+1}^{T} (y_t - \bar{y})(y_{t-\nu} - \bar{y})$.

However, for small samples, Eq. (1.5) does not necessarily result in a positive definite matrix. Newey and West (1987) have proposed an estimator that does not have this defect. Their estimator is given by:

$$S^{NW} = \hat{\gamma}_0 + 2 \sum_{\nu=1}^{q} \left(1 - \frac{\nu}{q+1}\right) \hat{\gamma}_\nu. \tag{1.6}$$

Loosely speaking, Theorem 1.1 says that, for a given sample size, the higher the "persistency" of a process, the lower the accuracy of the sample mean as an estimate of the population mean. To illustrate, consider three processes that feature the same marginal variance (equal to one, say), but different autocorrelations: 0%, 70%, and 99.9%. Figure 1.5 displays simulated paths of such three processes. It indeed appears that, the larger the autocorrelation of the process, the further the sample mean (dashed red line) from the population mean (red solid line).

The same type of simulations can be performed using this ShinyApp (use panel "AR(1)").

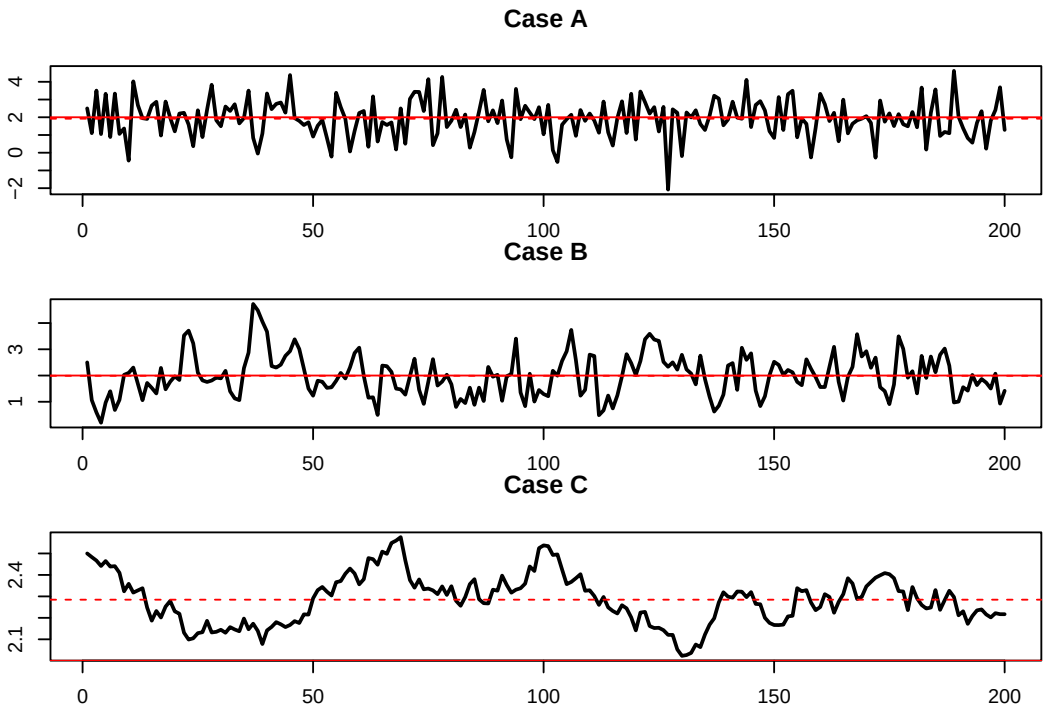


Figure 1.5: The three samples have been simulated using the following data generating process: $x_t = \mu + \rho(x_{t-1} - \mu) + \sqrt{1-\rho^2}\varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0,1)$. Case A: $\rho = 0$; Case B: $\rho = 0.7$; Case C: $\rho = 0.999$. In the three cases, $\mathbb{E}(x_t) = \mu = 2$ and $\mathbb{V}ar(x_t) = 1$.

Chapter 2

Univariate processes

2.0.1 Moving Average (MA) processes

Definition 2.1. Consider a white noise process $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ (Def. 1.1). Then y_t is a first-order moving average process if, for all t:

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}.$$

If $\mathbb{E}(\varepsilon_t^2) = \sigma^2$, it is easily obtained that the unconditional mean and variances of y_t are:

$$\mathbb{E}(y_t) = \mu, \quad \mathbb{V}ar(y_t) = (1+\theta^2)\sigma^2.$$

The first auto-covariance is:

$$\gamma_1 = \mathbb{E}\{(y_t - \mu)(y_{t-1} - \mu)\} = \theta\sigma^2.$$

Higher-order auto-covariances are zero ($\gamma_j = 0$ for $j > 1$). Therefore: An MA(1) process is covariance-stationary (Def. 1.4).

For a MA(1) process, the autocorrelation of order j (see Def. 1.6) is given by:

$$\rho_j = \begin{cases} 1 & \text{if} \quad j = 0, \\ \theta/(1+\theta^2) & \text{if} \quad j = 1 \\ 0 & \text{if} \quad j > 1. \end{cases}$$

Notice that process y_t defined through:

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1},$$

where $\mathbb{V}ar(\varepsilon_t) = \sigma^2$, has the same mean and autocovariances as

$$y_t = \mu + \varepsilon_t^* + \frac{1}{\theta}\varepsilon_{t-1}^*,$$

where $\mathbb{V}ar(\varepsilon_t^*) = \theta^2\sigma^2$. That is, even if we perfectly know the mean and autocovariances of this process, it is not possible to identify which specification is the one that has been used to generate the data. Only one of these two specifications is said to be *fundamental*, that is the one that satisfies $|\theta_1| < 1$ (see Eq. (2.26)).

Definition 2.2 (MA(q) process)**.** A q^{th} order Moving Average process is defined through:

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}.$$

where $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ is a white noise process (Def. 1.1).

Proposition 2.1 (Covariance-stationarity of an MA(q) process)**.** *Finite-order Moving Average processes are covariance-stationary.*

Moreover, the autocovariances of an MA(q) process (as defined in Def. 2.2) are given by:

$$\gamma_j = \begin{cases} \sigma^2(\theta_j\theta_0 + \theta_{j+1}\theta_1 + \cdots + \theta_q\theta_{q-j}) & \text{for} \quad j \in \{0, \dots, q\} \\ 0 & \text{for} \quad j > q, \end{cases} \tag{2.1}$$

where we use the notation $\theta_0 = 1$, and $\mathbb{V}ar(\varepsilon_t) = \sigma^2$.

Proof. The unconditional expectation of y_t does not depend on time, since $\mathbb{E}(y_t) = \mu$. Let's turn to autocovariances. We can extend the series of the θ_j's by setting $\theta_j = 0$ for $j > q$. We then have:

$$\begin{aligned} \mathbb{E}((y_t - \mu)(y_{t-j} - \mu)) &= \mathbb{E}\left[(\theta_0\varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_j\varepsilon_{t-j} + \theta_{j+1}\varepsilon_{t-j-1} + \dots) \times \right. \\ &\quad \left. (\theta_0\varepsilon_{t-j} + \theta_1\varepsilon_{t-j-1} + \dots)\right]. \end{aligned}$$

Then use the fact that $\mathbb{E}(\varepsilon_t\varepsilon_s) = 0$ if $t \neq s$ (because $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ is a white noise process). □

Figure 2.1 displays simulated paths of two MA processes (an MA(1) and an MA(4)). Such simulations can be produced by using panel "ARMA(p,q)" of this web interface.

```
library(AEC)
T <- 100;nb.sim <- 1
y.0 <- c(0)
c <- 1;phi <- c(0);sigma <- 1
theta <- c(1,1) # MA(1) specification
y.sim <- sim.arma(c,phi,theta,sigma,T,y.0,nb.sim)
par(mfrow=c(1,2))
par(plt=c(.2,.9,.2,.85))
plot(y.sim[,1],xlab="",ylab="",type="l",lwd=2,
     main=expression(paste(theta[0],"=1, ",theta[1],"=1",sep="")))
abline(h=c)
theta <- c(1,1,1,1,1) # MA(4) specification
y.sim <- sim.arma(c,phi,theta,sigma,T,y.0,nb.sim)
plot(y.sim[,1],xlab="",ylab="",type="l",lwd=2,
     main=expression(paste(theta[0],"=...=",theta[4],"=1",sep="")))
abline(h=c)
```

What if the order q of an MA(q) process gets infinite? The notion of **infinite-order Moving Average process** exists and is important in time series analysis. The (infinite) sequence of θ_j has to satisfy some conditions for such a process to be well-defined (see Theorem 2.1 below). These conditions relate to the "summability" of $\{\theta_i\}_{i\in\mathbb{N}}$ (see Definition 2.3).

Definition 2.3 (Absolute and square summability)**.** The sequence $\{\theta_i\}_{i\in\mathbb{N}}$ is absolutely summable if $\sum_{i=0}^{\infty}|\theta_i| < +\infty$, and it is square summable if $\sum_{i=0}^{\infty}\theta_i^2 < +\infty$.

According to Prop. 8.8, absolute summability implies square summability.

Theorem 2.1 (Existence condition for an infinite MA process)**.** *If $\{\theta_i\}_{i\in\mathbb{N}}$ is square summable (see Def. 2.3) and if $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ is a white noise process (see Def. 1.1), then*

$$\mu + \sum_{i=0}^{+\infty}\theta_i\varepsilon_{t-i}$$

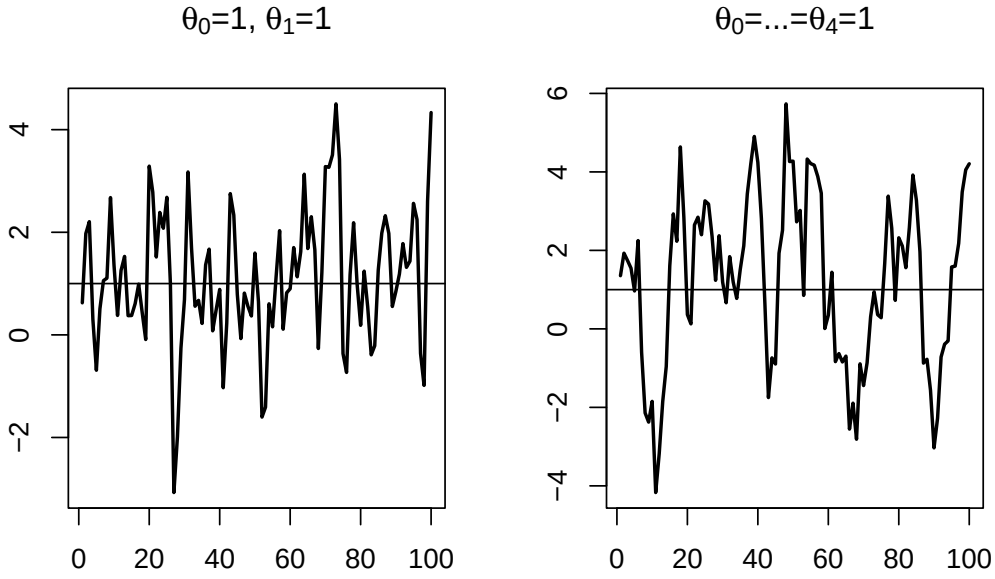


Figure 2.1: Simulation of MA processes.

defines a well-behaved [covariance-stationary] process, called infinite-order MA process (MA(∞)).

Proof. See Appendix 3.A in Hamilton. "Well behaved" means that $\Sigma_{i=0}^{T}\theta_{t-i}\varepsilon_{t-i}$ converges in mean square (Def. 8.17) to some random variable Z_t. The proof makes use of the fact that:

$$\mathbb{E}\left[\left(\sum_{i=N}^{M}\theta_i\varepsilon_{t-i}\right)^2\right] = \sum_{i=N}^{M}|\theta_i|^2\sigma^2,$$

and that, when $\{\theta_i\}$ is square summable, $\forall\eta > 0$, $\exists N$ s.t. the right-hand-side term in the last equation is lower than η for all $M \geq N$ (static Cauchy criterion, Theorem 8.2). This implies that $\Sigma_{i=0}^{T}\theta_i\varepsilon_{t-i}$ converges in mean square (stochastic Cauchy criterion, see Theorem 8.3). □

Proposition 2.2 (First two moments of an infinite MA process)**.** *If $\{\theta_i\}_{i\in\mathbb{N}}$ is absolutely summable, i.e. if $\sum_{i=0}^{\infty}|\theta_i| < +\infty$, then*

i. $y_t = \mu + \sum_{i=0}^{+\infty} \theta_i \varepsilon_{t-i}$ *exists (Theorem 2.1) and is such that:*

$$\begin{aligned} \mathbb{E}(y_t) &= \mu \\ \gamma_0 = \mathbb{E}([y_t - \mu]^2) &= \sigma^2(\theta_0^2 + \theta_1^2 + \dots) \\ \gamma_j = \mathbb{E}([y_t - \mu][y_{t-j} - \mu]) &= \sigma^2(\theta_0\theta_j + \theta_1\theta_{j+1} + \dots). \end{aligned}$$

ii. Process y_t *has absolutely summable auto-covariances, which implies that the results of Theorem 1.1 (Central Limit) apply.*

Proof. The absolute summability of $\{\theta_i\}$ and the fact that $\mathbb{E}(\varepsilon^2) < \infty$ imply that the order of integration and summation is interchangeable (see Hamilton, 1994, Footnote p. 52), which proves (i). For (ii), see end of Appendix 3.A in Hamilton (1994). □

2.0.2 Auto-Regressive (AR) processes

Definition 2.4 (First-order AR process (AR(1)))**.** Consider a white noise process $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ (see Def. 1.1). Process y_t is an AR(1) process if it is defined by the following difference equation:

$$y_t = c + \phi y_{t-1} + \varepsilon_t.$$

If $|\phi| \geq 1$, y_t is not stationary. Indeed, we have:

$$y_{t+k} = c + \varepsilon_{t+k} + \phi(c + \varepsilon_{t+k-1}) + \phi^2(c + \varepsilon_{t+k-2}) + \cdots + \phi^{k-1}(c + \varepsilon_{t+1}) + \phi^k y_t.$$

Therefore, the conditional variance

$$\mathbb{V}ar_t(y_{t+k}) = \sigma^2(1 + \phi^2 + \phi^4 + \cdots + \phi^{2(k-1)})$$

does not converge for large k's. This implies that $\mathbb{V}ar(y_t)$ does not exist.

By contrast, if $|\phi| < 1$, one can see that:

$$y_t = c + \varepsilon_t + \phi(c + \varepsilon_{t-1}) + \phi^2(c + \varepsilon_{t-2}) + \cdots + \phi^k(c + \varepsilon_{t-k}) + \dots$$

Hence, if $|\phi| < 1$, the unconditional mean and variance of y_t are:

$$\mathbb{E}(y_t) = \frac{c}{1-\phi} =: \mu \quad \text{and} \quad \mathbb{V}ar(y_t) = \frac{\sigma^2}{1-\phi^2}.$$

Let us compute the j^{th} autocovariance of the AR(1) process:

$$\begin{aligned}\mathbb{E}([y_t-\mu][y_{t-j}-\mu]) &= \mathbb{E}([\varepsilon_t+\phi\varepsilon_{t-1}+\phi^2\varepsilon_{t-2}+\cdots+\phi^j\varepsilon_{t-j}+\phi^{j+1}\varepsilon_{t-j-1}\ldots]\times \\ &\quad [\varepsilon_{t-j}+\phi\varepsilon_{t-j-1}+\phi^2\varepsilon_{t-j-2}+\cdots+\phi^k\varepsilon_{t-j-k}+\ldots]) \\ &= \mathbb{E}(\phi^j\varepsilon_{t-j}^2+\phi^{j+2}\varepsilon_{t-j-1}^2+\phi^{j+4}\varepsilon_{t-j-2}^2+\ldots) \\ &= \frac{\phi^j\sigma^2}{1-\phi^2}.\end{aligned}$$

Therefore $\rho_j=\phi^j$.

By what precedes, we have:

Proposition 2.3 (Covariance-stationarity of an AR(1) process)**.** *The AR(1) process, as defined in Def. 2.4, is covariance-stationary iff* $|\phi|<1$.

Definition 2.5 (AR(p) process)**.** Consider a white noise process $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ (see Def. 1.1). Process y_t is a p^{th}-order autoregressive process (AR(p)) if its dynamics is defined by the following difference equation (with $\phi_p\neq 0$):

$$y_t=c+\phi_1y_{t-1}+\phi_2y_{t-2}+\cdots+\phi_py_{t-p}+\varepsilon_t. \tag{2.2}$$

As we will see, the covariance-stationarity of process y_t hinges on matrix F defined as:

$$F=\begin{bmatrix}\phi_1 & \phi_2 & \ldots & & \phi_p\\ 1 & 0 & \ldots & & 0\\ 0 & 1 & \ldots & & 0\\ \vdots & & \ddots & & \vdots\\ 0 & 0 & \ldots & 1 & 0\end{bmatrix}. \tag{2.3}$$

Note that this matrix F is such that if y_t follows Eq. (2.2), then process $\mathbf{y}_t$ follows:

$$\mathbf{y}_t=\mathbf{c}+F\mathbf{y}_{t-1}+\xi_t$$

with

$$\mathbf{c}=\begin{bmatrix}c\\0\\\vdots\\0\end{bmatrix},\quad \xi_t=\begin{bmatrix}\varepsilon_t\\0\\\vdots\\0\end{bmatrix},\quad \mathbf{y}_t=\begin{bmatrix}y_t\\y_{t-1}\\\vdots\\y_{t-p+1}\end{bmatrix}.$$

Proposition 2.4 (The eigenvalues of matrix F)**.** *The eigenvalues of F (defined by Eq.* (2.3)*) are the solutions of:*

$$\lambda^p - \phi_1 \lambda^{p-1} - \cdots - \phi_{p-1}\lambda - \phi_p = 0. \tag{2.4}$$

Proposition 2.5 (Covariance-stationarity of an AR(p) process)**.** *These four statements are equivalent:*

i. Process $\{y_t\}$, defined in Def. 2.5, is covariance-stationary.
ii. The eigenvalues of F (as defined Eq. (2.3)*) lie strictly within the unit circle.*
iii. The roots of Eq. (2.5) *(below) lie strictly outside the unit circle.*

$$1 - \phi_1 z - \cdots - \phi_{p-1} z^{p-1} - \phi_p z^p = 0. \tag{2.5}$$

iv. The roots of Eq. (2.6) *(below) lie strictly inside the unit circle.*

$$\lambda^p - \phi_1 \lambda^{p-1} - \cdots - \phi_{p-1}\lambda - \phi_p = 0. \tag{2.6}$$

Proof. We consider the case where the eigenvalues of F are distinct; Jordan decomposition can be used in the general case. When the eigenvalues of F are distinct, F admits the following spectral decomposition: $F = PDP^{-1}$, where D is diagonal. Using the notations introduced in Eq. (2.3), we have:

$$\mathbf{y}_t = \mathbf{c} + F\mathbf{y}_{t-1} + \xi_t.$$

Let's introduce $\mathbf{d} = P^{-1}\mathbf{c}$, $\mathbf{z}_t = P^{-1}\mathbf{y}_t$ and $\eta_t = P^{-1}\xi_t$. We have:

$$\mathbf{z}_t = \mathbf{d} + D\mathbf{z}_{t-1} + \eta_t.$$

Because D is diagonal, the different component of $\mathbf{z}_t$, denoted by $z_{i,t}$, follow AR(1) processes. The (scalar) autoregressive parameters of these AR(1) processes are the diagonal entries of D –which also are the eigenvalues of F– that we denote by λ_i.

Process y_t is covariance-stationary iff $\mathbf{y}_t$ also is covariance-stationary, which is the case iff all $z_{i,t}$, $i \in [1, p]$, are covariance-stationary. By Prop. 2.3, process $z_{i,t}$ is covariance-stationary iff $|\lambda_i| < 1$. This proves that (i) is equivalent to (ii). Prop. 2.4 further proves that (ii) is equivalent to (iv). Finally, it is easily seen that (iii) is equivalent to (iv) (as long as $\phi_p \neq 0$). □

Using the lag operator (see Eq (1.1)), if y_t is a covariance-stationary AR(p) process (Def. 2.5), we can write:

$$y_t = \mu + \psi(L)\varepsilon_t,$$

where

$$\psi(L) = (1 - \phi_1 L - \cdots - \phi_p L^p)^{-1}, \tag{2.7}$$

and

$$\mu = \mathbb{E}(y_t) = \frac{c}{1 - \phi_1 - \cdots - \phi_p}. \tag{2.8}$$

In the following lines of codes, we compute the eigenvalues of the F matrices associated with the following processes (where ε_t is a white noise):

$$\begin{aligned} x_t &= 0.9x_{t-1} - 0.2x_{t-2} + \varepsilon_t \\ y_t &= 1.1y_{t-1} - 0.3y_{t-2} + \varepsilon_t \\ w_t &= 1.4w_{t-1} - 0.7w_{t-2} + \varepsilon_t \\ z_t &= 0.9z_{t-1} + 0.2z_{t-2} + \varepsilon_t \end{aligned}$$

```
F <- matrix(c(.9,1,-.2,0),2,2)
lambda_x <- eigen(F)$values
F[1,] <- c(1.1,-.3)
lambda_y <- eigen(F)$values
F[1,] <- c(1.4,-.7)
lambda_w <- eigen(F)$values
F[1,] <- c(.9,.2)
lambda_z <- eigen(F)$values
rbind(lambda_x,lambda_y,lambda_w,lambda_z)
```

```
##                          [,1]                  [,2]
## lambda_x 0.500000+0.0000000i  0.4000000+0.0000000i
## lambda_y 0.600000+0.0000000i  0.5000000+0.0000000i
## lambda_w 0.700000+0.4582576i  0.7000000-0.4582576i
## lambda_z 1.084429+0.0000000i -0.1844289+0.0000000i
```

The absolute values of the eigenvalues associated with process w_t are both equal to 0.837. Therefore, according to Proposition 2.5, processes x_t, y_t, and

w_t are covariance-stationary, but not z_t (because the absolute value of one of the eigenvalues of the F matrix associated with this process is larger than 1).

The computation of the autocovariances of y_t is based on the so-called **Yule-Walker equations** (Eq. (2.9)). Let's rewrite Eq. (2.2):

$$(y_t - \mu) = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \cdots + \phi_p(y_{t-p} - \mu) + \varepsilon_t.$$

Multiplying both sides by $y_{t-j} - \mu$ and taking expectations leads to the (Yule-Walker) equations:

$$\gamma_j = \begin{cases} \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \cdots + \phi_p\gamma_{j-p} & if \quad j > 0 \\ \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p + \sigma^2 & for \quad j = 0. \end{cases} \tag{2.9}$$

Using $\gamma_j = \gamma_{-j}$ (Prop. 1.1), one can express $(\gamma_0, \gamma_1, \ldots, \gamma_p)$ as functions of $(\sigma^2, \phi_1, \ldots, \phi_p)$.

2.0.3 AR-MA processes

Definition 2.6 (ARMA(p,q) process)**.** $\{y_t\}$ is an ARMA(p,q) process if its dynamics is described by the following equation:

$$y_t = c + \underbrace{\phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}}_{\text{AR part}} + \underbrace{\varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}}_{\text{MA part}}, \tag{2.10}$$

where $\{\varepsilon_t\}_{t\in]-\infty,+\infty[}$ is a white noise process (see Def. 1.1).

Proposition 2.6 (Stationarity of an ARMA(p,q) process)**.** *The ARMA(p,q) process defined in 2.6 is covariance stationary iff the roots of*

$$1 - \phi_1 z - \cdots - \phi_p z^p = 0$$

lie strictly outside the unit circle or, equivalently, iff those of

$$\lambda^p - \phi_1\lambda^{p-1} - \cdots - \phi_p = 0$$

lie strictly within the unit circle.

Proof. The proof of Prop. 2.5 can be adapted to the present case. □

We can write:

$$(1 - \phi_1 L - \cdots - \phi_p L^p) y_t = c + (1 + \theta_1 L + \cdots + \theta_q L^q)\varepsilon_t.$$

If the roots of $1 - \phi_1 z - \cdots - \phi_p z^p = 0$ lie outside the unit circle, we have:

$$y_t = \mu + \psi(L)\varepsilon_t, \tag{2.11}$$

where

$$\psi(L) = \frac{1 + \theta_1 L + \cdots + \theta_q L^q}{1 - \phi_1 L - \cdots - \phi_p L^p} \quad and \quad \mu = \frac{c}{1 - \phi_1 - \cdots - \phi_p}.$$

Eq. (2.11) is the **Wold representation** of this ARMA process (see Theorem 2.2 below).

The stationarity of the process depends only on the AR specification (or on the eigenvalues of matrix F, exactly as in Prop. 2.5). If the process is stationary, the weights in $\psi(L)$ decay at a geometric rate.

2.0.4 PACF approach to identify AR/MA processes

We have seen that the k^{th}-order auto-correlation of a MA(q) process is null if $k > q$. This is exploited, in practice, to determine the order of a MA process. Moreover, since this is not the case for an AR process, this can be used to distinguish an AR from an MA process.

There exists an equivalent approach to determine whether a process can be modeled as an AR process; it is based on partial auto-correlations:

Definition 2.7 (Partial auto-correlation)**.** In a time series context, the partial auto-correlation ($\phi_{h,h}$) of process $\{y_t\}$ is defined as the partial correlation of y_{t+h} and y_t given $y_{t+h-1}, \ldots, y_{t+1}$. (see Def. 8.5 for the definition of partial correlation.)

If $h > p$, the regression of y_{t+h} on $y_{t+h-1}, \ldots, y_{t+1}$ is:

$$y_{t+h} = c + \phi_1 y_{t+h-1} + \cdots + \phi_p y_{t+h-p} + \varepsilon_{t+h}.$$

The residuals of the latter regressions (ε_{t+h}) are uncorrelated to y_t. Then the partial autocorrelation is zero for $h > p$.

Besides, it can be shown that $\phi_{p,p} = \phi_p$. Hence $\phi_{p,p} = \phi_p$ but $\phi_{h,h} = 0$ for $h > p$. This can be used to determine the order of an AR process. By contrast (importantly) if y_t follows an MA(q) process, then $\phi_{k,k}$ asymptotically approaches zero instead of cutting off abruptly.

As illustrated below, functions `acf` and `pacf` can be conveniently used to employ the (P)ACF approach. (Note also the use of function `sim.arma` to simulate ARMA processes.)

```
library(AEC)
par(mfrow=c(3,2))
par(plt=c(.2,.9,.2,.95))
theta <- c(1,2,1);phi=0
y.sim <- sim.arma(c=0,phi,theta,sigma=1,T=1000,y.0=0,nb.sim=1)
par(mfg=c(1,1));plot(y.sim,type="l",lwd=2)
par(mfg=c(2,1));acf(y.sim)
par(mfg=c(3,1));pacf(y.sim)
theta <- c(1);phi=0.9
y.sim <- sim.arma(c=0,phi,theta,sigma=1,T=1000,y.0=0,nb.sim=1)
par(mfg=c(1,2));plot(y.sim,type="l",lwd=2)
par(mfg=c(2,2));acf(y.sim)
par(mfg=c(3,2));pacf(y.sim)
```

2.0.5 Wold decomposition

The Wold decomposition is an important result in time series analysis:

Theorem 2.2 (Wold decomposition)**.** *Any covariance-stationary process admits the following representation:*

$$y_t = \mu + \sum_{0}^{+\infty} \theta_i \varepsilon_{t-i} + \kappa_t,$$

where

- $\theta_0 = 1$, $\sum_{i=0}^{\infty} \theta_i^2 < +\infty$ *(square summability, see Def. 2.3).*

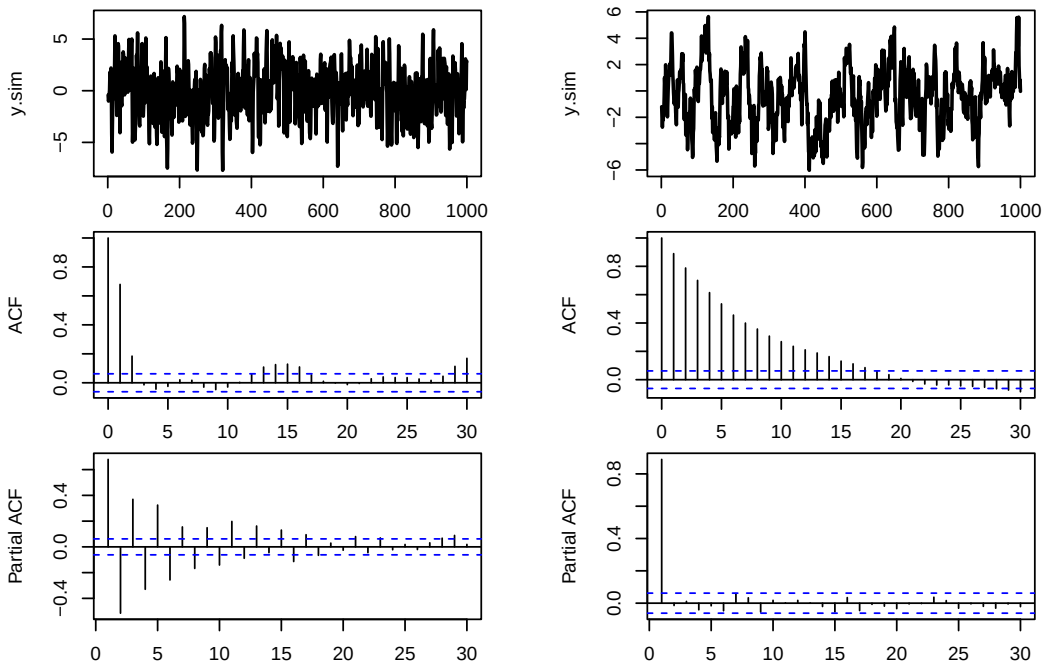


Figure 2.2: ACF/PACF analysis of two processes (MA process on the left, AR on the right).

- *$\{\varepsilon_t\}$ is a white noise (see Def. 1.1); ε_t is the error made when forecasting y_t based on a linear combination of lagged y_t's ($\varepsilon_t = y_t - \widehat{\mathbb{E}}[y_t|y_{t-1}, y_{t-2}, ...]$).*
- *For any $j \geq 1$, κ_t is not correlated with ε_{t-j}; but κ_t can be perfectly forecasted based on a linear combination of lagged y_t's (i.e. $\kappa_t = \widehat{\mathbb{E}}(\kappa_t|y_{t-1}, y_{t-2}, ...)$). κ_t is called the **deterministic component** of y_t.*

Proof. See Anderson (1971). Partial proof in L. Christiano. □

For an ARMA process, the Wold representation is given by Eq. (2.11). As detailed in Prop. 2.7, it can be computed by recursively replacing the lagged y_t's in Eq. (2.10). In this case, the deterministic component (κ) is null.

2.0.6 Impulse Response Functions (IRFs) in ARMA models

Consider the ARMA(p,q) process defined in Def. 2.6, whose associated sequence of white noise is $\{\varepsilon_t\}$. Let us construct a novel (counterfactual) sequence of shocks $\{\tilde{\varepsilon}_t^{(s)}\}$:

$$\tilde{\varepsilon}_t^{(s)} = \begin{cases} \varepsilon_t & if \quad t \neq s, \\ \varepsilon_t + \delta & if \quad t = s. \end{cases}$$

We denote by $\{\tilde{y}_t^{(s)}\}$ the process following Eq. (2.10) where $\{\varepsilon_t\}$ is replaced with $\{\tilde{\varepsilon}_t^{(s)}\}$. The time series $\{\tilde{y}_t^{(s)}\}$ is the counterfactual series $\{y_t\}$ that would have prevailed if ε_t had been shifted by δ on date s (and that would be the only change).

The relationship between $\{y_t\}$ and $\{\tilde{y}_t^{(s)}\}$ defines the **dynamic multiplier**. The latter is denoted by $\frac{\partial y_t}{\partial \varepsilon_s}$ and is such that:

$$\tilde{y}_t^{(s)} = y_t + \frac{\partial y_t}{\partial \varepsilon_s}\delta.$$

We will see that the dynamic multipliers are closely related to the infinite MA representation (or **Wold decomposition**, Theorem 2.2) of y_t:

$$y_t = \mu + \sum_{i=0}^{+\infty} \psi_i \varepsilon_{t-i}.$$

For $t < s$, we have $y_t = \tilde{y}_t^{(s)}$ (because $\tilde{\varepsilon}_{t-i} = \varepsilon_{t-i}$ for all $i \geq 0$ if $t < s$).

For $t \geq s$:

$$\tilde{y}_t^{(s)} = \mu + \left(\sum_{i=0}^{t-s-1} \psi_i \varepsilon_{t-i} \right) + \psi_{t-s}(\varepsilon_s + \delta) + \left(\sum_{i=t-s+1}^{+\infty} \psi_i \varepsilon_{t-i} \right) = y_t + \frac{\partial y_t}{\partial \varepsilon_s} \delta.$$

Therefore, for $t \geq s$, we have:

$$\boxed{\frac{\partial y_t}{\partial \varepsilon_s} = \psi_{t-s}.}$$

That is, $\{y_t\}$'s dynamic multiplier of order k is the same object as the k^{th} loading ψ_k in the Wold decomposition of $\{y_t\}$. The sequence $\left\{ \frac{\partial y_{t+h}}{\partial \varepsilon_t} \right\}_{h \geq 0} \equiv \{\psi_h\}_{h \geq 0}$ defines the **impulse response function (IRF)** of y_t to the shock ε_t.

For ARMA processes, the computation of the IRFs is easy:

Proposition 2.7 (IRF of an ARMA(p,q) process)**.** *The coefficients ψ_h, that define the IRF of process y_t to ε_t, can be computed recursively as follows:*

1. *Set $\psi_{-1} = \cdots = \psi_{-p} = 0$.*
2. *For $h \geq 0$, (recursively) apply:*

$$\psi_h = \phi_1 \psi_{h-1} + \cdots + \phi_p \psi_{h-p} + \theta_h,$$

where $\theta_h = 0$ for $h > q$.

Proof. This is obtained by applying the operator $\frac{\partial}{\partial \varepsilon_t}$ on both sides of Eq. (2.10):

$$y_{t+h} = c + \phi_1 y_{t+h-1} + \cdots + \phi_p y_{t+h-p} + \varepsilon_{t+h} + \theta_1 \varepsilon_{t+h-1} + \cdots + \theta_q \varepsilon_{t+h-q}.$$

□

Note that Proposition 2.7 constitutes a simple way to compute the MA(∞) representation (or Wold representation) of an ARMA process.

One can use function `sim.arma` of package `AEC` to compute ARMA's IRFs (with the argument `make.IRF = 1`):

```
T <- 21 # number of periods for IRF
theta <- c(1,1,1);phi <- c(0);c <- 0
y.sim <- sim.arma(c,phi,theta,sigma=1,T,y.0=rep(0,length(phi)),
                  nb.sim=1,make.IRF = 1)
par(mfrow=c(1,3));par(plt=c(.25,.95,.2,.85))
plot(0:(T-1),y.sim[,1],type="l",lwd=2,
     main="(a) Process 1",xlab="Time after shock on epsilon",
     ylab="Dynamic multiplier (shock on epsilon at t=0)",col="red")
abline(h=0)
theta <- c(1,.5);phi <- c(0.6)
y.sim <- sim.arma(c,phi,theta,sigma=1,T,y.0=rep(0,length(phi)),
                  nb.sim=1,make.IRF = 1)
plot(0:(T-1),y.sim[,1],type="l",lwd=2,
     main="(b) Process 2",xlab="Time after shock on epsilon",
     ylab="",col="red")
theta <- c(1,1,1);phi <- c(0,0,.5,.4)
y.sim <- sim.arma(c,phi,theta,sigma=1,T,y.0=rep(0,length(phi)),
                  nb.sim=1,make.IRF = 1)
plot(0:(T-1),y.sim[,1],type="l",lwd=2,
     main="(c) Process 3",xlab="Time after shock on epsilon",
     ylab="",col="red")
```

Consider the annual Swiss GDP growth from the JST macro-history database. Let us first determine relevant orders for AR and MA processes using the (P)ACF approach.

```
library(AEC)
data(JST)
data <- subset(JST,iso=="CHE")
par(plt=c(.1,.95,.1,.95))
T <- dim(data)[1]
growth <- log(data$gdp[2:T]/data$gdp[1:(T-1)])
par(mfrow=c(3,1))
par(plt=c(.1,.95,.15,.95))
plot(data$year[2:T],growth,type="l",xlab="",ylab="",lwd=2)
abline(h=0,lty=2)
acf(growth)
```

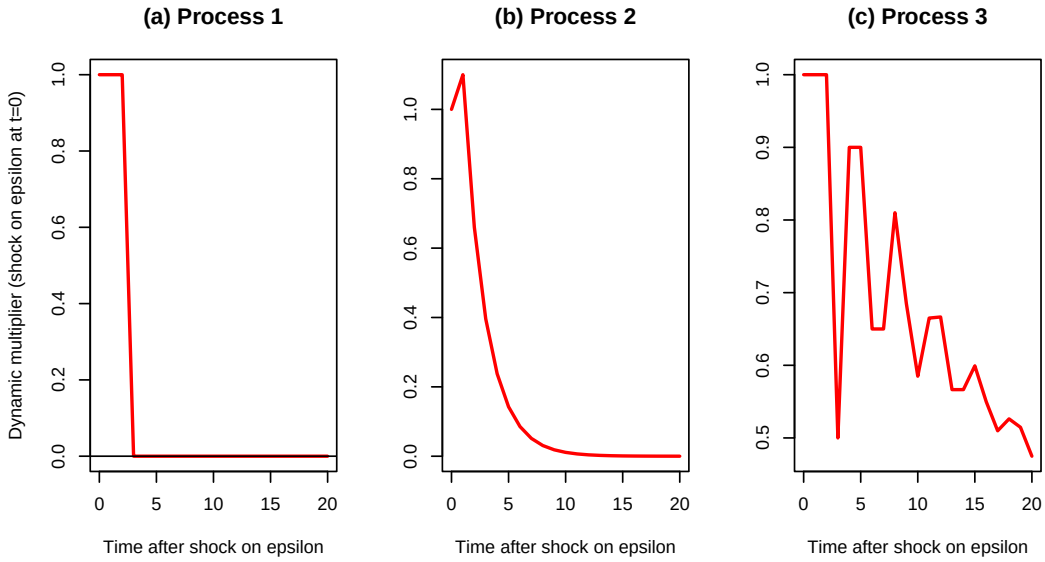


Figure 2.3: IRFs associated with the three processes. Process 1 (MA(2)): $y_t = \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2}$. Process 2 (ARMA(1,1)): $y_t = 0.6y_{t-1} + \varepsilon_t + 0.5\varepsilon_{t-1}$. Process 3 (ARMA(4,2)): $y_t = 0.5y_{t-3} + 0.4y_{t-4} + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2}$.

```
pacf(growth)
```

The two bottom plots of Figure 2.4 suggest that either an MA(2) or an AR(1) could be used to model the GDP growth rate series. Figure 2.5 shows the IRFs based on these two respective specifications.

```
# Fit an AR process:
res <- arima(growth,order=c(1,0,0))
phi <- res$coef[1]
T <- 11
y.sim <- sim.arma(c=0,phi,theta=1,sigma=1,T,y.0=rep(0,length(phi)),
                  nb.sim=1,make.IRF = 1)
par(plt=c(.15,.95,.25,.95))
plot(0:(T-1),y.sim[,1],type="l",lwd=3,
     xlab="Time after shock on epsilon",
     ylab="Dynamic multiplier (shock on epsilon at t=0)",col="red")
# Fit a MA process:
```

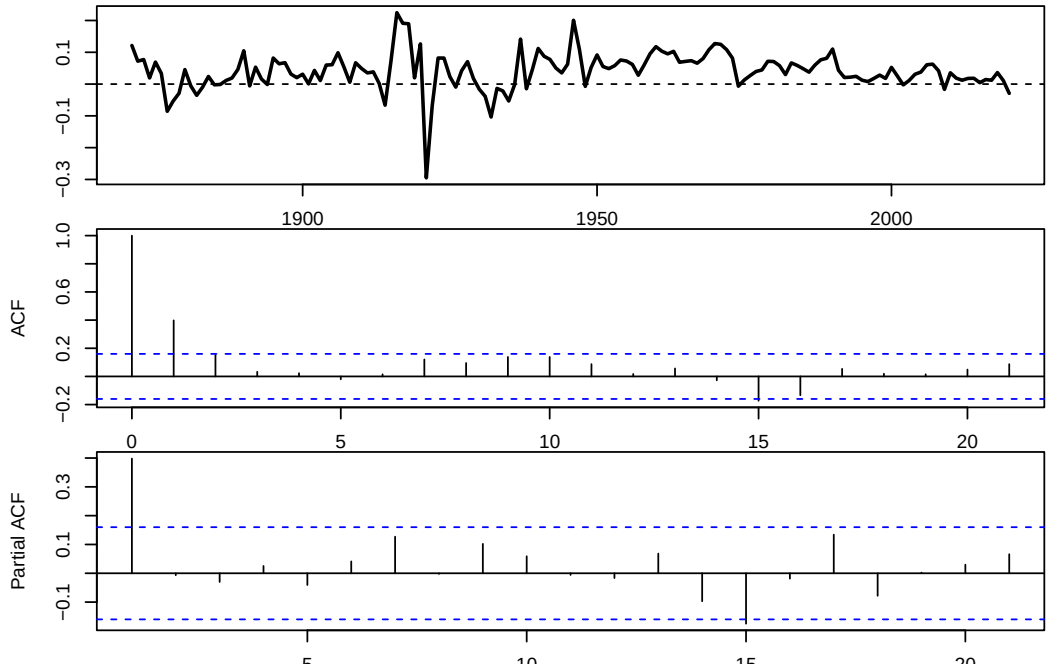


Figure 2.4: (P)ACF analysis of Swiss GDP growth.

```
res <- arima(growth,order=c(0,0,2))
phi <- 0;theta <- c(1,res$coef[1:2])
y.sim <- sim.arma(c=0,phi,theta,sigma=1,T,y.0=rep(0,length(phi)),
                  nb.sim=1,make.IRF = 1)
lines(0:(T-1),y.sim[,1],lwd=3,col="red",lty=2)
abline(h=0)
```

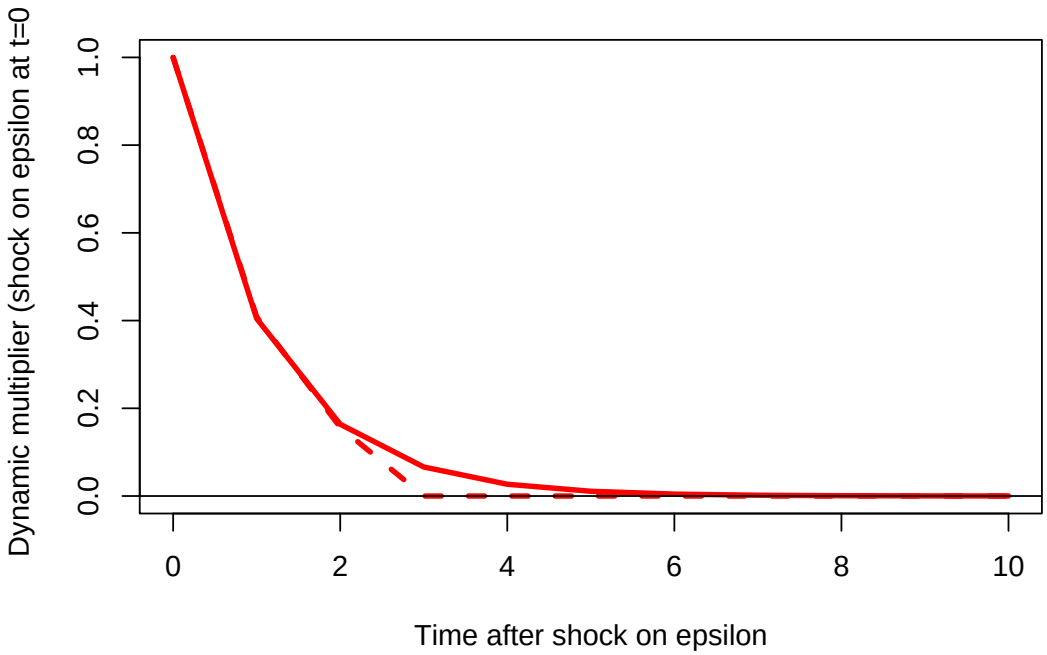


Figure 2.5: Dynamic response of Swiss annual growth to a shock on the innovation ε_t at date $t = 0$. The solid line corresponds to an AR(1) specification; the dashed line corresponds to a MA(2) specification.

The same kind of algorithm can be used to compute the impact of an increase in an exogenous variable x_t within an ARMAX(p,q,r) model (see next section).

2.0.7 ARMA processes with exogenous variables (ARMA-X)

ARMA processes do not allow to investigate the influence of an exogenous variable (say x_t) on the variable of interest (say y_t). When x_t and y_t have reciprocal influences, the Vector Autoregressive (VAR) model may be used (this tools will be studied later, in Section 3). However, when one suspects that x_t has an "exogenous" influence on y_t, then a simple extension of the ARMA processes may be considered. Loosely speaking, x_t has an "exogenous" influence on y_t if y_t does not affect x_t. This extension is called ARMAX(p,q,r).

To begin with, let us formalize this notion of exogeneity. Consider a white noise sequence $\{\varepsilon_t\}$ (Def. 1.1).

Definition 2.8 (Exogeneity)**.** We say that x_t is (strictly) exogenous to $\{\varepsilon_t\}$ if

$$\mathbb{E}(\varepsilon_t| \underbrace{..., x_{t+1}}_{\text{future}}, \quad \underbrace{x_t, x_{t-1}, ...}_{\text{present and past}}) = 0.$$

Hence, if $\{x_t\}$ is strictly exogenous to ε_t, then past, present and future values of x_t do not allow to predict the ε_t's.

In the following, we assume that $\{x_t\}$ is a covariance stationary process.

Definition 2.9 (ARMAX(p,q,r) model)**.** The process $\{y_t\}$ is an ARMAX(p,q,r) if it follows a difference equation:

$$\begin{aligned} y_t \;=\; & \underbrace{c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}}_{\text{AR(p) part}} + \underbrace{\beta_0 x_t + \cdots + \beta_r x_{t-r}}_{\text{X(r) part}} + \\ & \underbrace{\varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.}_{\text{MA(q) part}} \end{aligned} \tag{2.12}$$

where $\{\varepsilon_t\}$ is an i.i.d. white noise sequence and $\{x_t\}$ is exogenous to y_t.

What is the effect of a one-unit increase in x_t on y_t? To address this question, this notion of "effect" has to be formalized. Let us introduce two related sequences of values for $\{x\}$. Denote the first by $\{a\}$ and the second by $\{\tilde{a}^t\}$. Further, we posit $a_s = \tilde{a}^t_s$ for all $s \neq t$, and $\tilde{a}^t_t = a_t + 1$.

With these notations, we define $\frac{\partial y_{t+h}}{\partial x_t}$ as follows:

$$\frac{\partial y_{t+h}}{\partial x_t} := \mathbb{E}(y_{t+h}|\{x\} = \{\tilde{a}^t\}) - \mathbb{E}(y_{t+h}|\{x\} = \{a\}). \tag{2.13}$$

Under the exogeneity assumption, it is easily seen that

$$\frac{\partial y_t}{\partial x_t} = \beta_0.$$

Now, since

$$\begin{aligned} y_{t+1} &= c + \phi_1 y_t + \cdots + \phi_p y_{t+1-p} + \beta_0 x_{t+1} + \cdots + \beta_r x_{t+1-r} + \\ &\quad \varepsilon_{t+1} + \theta_1 \varepsilon_t + \cdots + \theta_q \varepsilon_{t+1-q}, \end{aligned}$$

and using the exogeneity assumption, we obtain:

$$\frac{\partial y_{t+1}}{\partial x_t} := \phi_1 \frac{\partial y_t}{\partial x_t} + \beta_1 = \phi_1 \beta_0 + \beta_1.$$

This can be applied recursively to give $\frac{\partial y_{t+h}}{\partial x_t}$ for any $h \geq 0$:

Proposition 2.8 (Dynamic multipliers in ARMAX models)**.** *One can recursively compute the dynamic multipliers* $\frac{\partial y_{t+h}}{\partial x_t}$ *as follows:*

i. Initialization: $\frac{\partial y_{t+h}}{\partial x_t} = 0$ *for* $h < 0$.

ii. For $h \geq 0$ *and assuming that the first* $h-1$ *multipliers have been computed, we have:*

$$\frac{\partial y_{t+h}}{\partial x_t} = \phi_1 \frac{\partial y_{t+h-1}}{\partial x_t} + \cdots + \phi_p \frac{\partial y_{t+h-p}}{\partial x_t} + \beta_h, \tag{2.14}$$

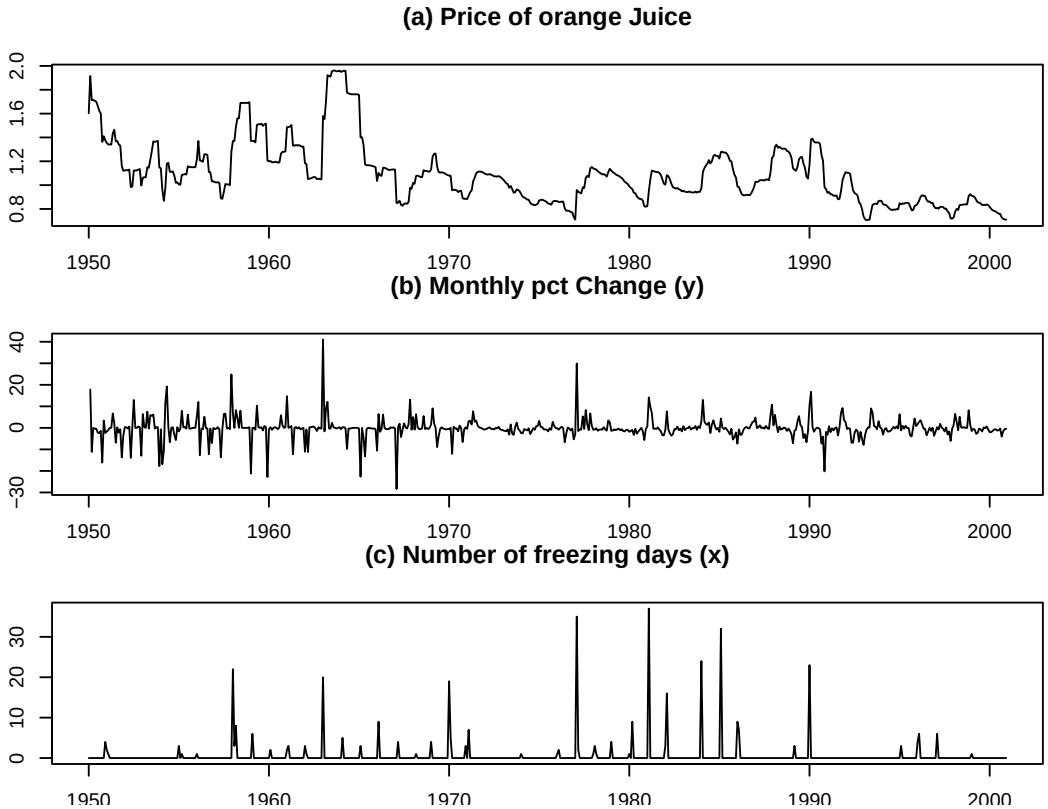
where we use the notation $\beta_h = 0$ *if* $h > r$.

Remark that the resulting dynamic multipliers are the same as those obtained for an ARMA(p,r) model where the θ_i's are replaced with β_i's (see Proposition 2.7 in Section 2.0.7).

It has to be stressed that the definition of the dynamic multipliers (Eq. (2.14)) does not reflect a potential persistency of the shock occuring on date t in process $\{x\}$ itself. Going in this direction would necessitate to model the joint dynamics of x_t (for instance using a VAR model , see Section 3).

Example 2.1 (Influence of the number of freezing days on the price of orange juice)**.** This example is based on data used in Stock and Watson (2003) (Chapter 16). The objective is to study the influence of the number of freezing days on the price of orange juice. Let us first estimate a ARMAX(0,0,12) model:

```
library(AEC)
library(AER)
data("FrozenJuice")
FJ <- as.data.frame(FrozenJuice)
date <- time(FrozenJuice)
price <- FJ$price/FJ$ppi
T <- length(price)
k <- 1
dprice <- 100*log(price[(k+1):T]/price[1:(T-k)])
fdd <- FJ$fdd[(k+1):T]
par(mfrow=c(3,1))
par(plt=c(.1,.95,.15,.75))
plot(date,price,type="l",xlab="",ylab="",
     main="(a) Price of orange Juice")
plot(date,c(NaN,dprice),type="l",xlab="",ylab="",
     main="(b) Monthly pct Change (y)")
plot(date,FJ$fdd,type="l",xlab="",ylab="",
     main="(c) Number of freezing days (x)")
```




```
nb.lags <- 12
FDD <- FJ$fdd[(nb.lags+1):T]
names.FDD <- NULL
for(i in 1:nb.lags){
  FDD <- cbind(FDD,FJ$fdd[(nb.lags+1-i):(T-i)])
  names.FDD <- c(names.FDD,paste(" Lag ",toString(i),sep=""))}
colnames(FDD) <- c(" Lag 0",names.FDD)
dprice <- dprice[(length(dprice)-dim(FDD)[1]+1):length(dprice)]
eq <- lm(dprice~FDD)
# Compute the Newey-West std errors:
var.cov.mat <- NeweyWest(eq,lag = 7, prewhite = FALSE)
robust_se <- sqrt(diag(var.cov.mat))
# Stargazer output (with and without Robust SE)
stargazer::stargazer(eq, eq, type = "text",
                     column.labels=c("(no HAC)","(HAC)"),keep.stat="n",
                     se = list(NULL,robust_se),no.space = TRUE)
```

```
##
## ========================================
##                  Dependent variable:
##              ---------------------------
##                         dprice
##                  (no HAC)        (HAC)
##                    (1)             (2)
## ----------------------------------------
## FDD Lag 0       0.496***       0.496***
##                 (0.058)         (0.139)
## FDD Lag 1       0.150***        0.150*
##                 (0.058)         (0.087)
## FDD Lag 2        0.046           0.046
##                 (0.057)         (0.056)
## FDD Lag 3        0.062           0.062
##                 (0.057)         (0.046)
## FDD Lag 4        0.024           0.024
##                 (0.057)         (0.030)
## FDD Lag 5        0.036           0.036
##                 (0.057)         (0.030)
## FDD Lag 6        0.037           0.037
##                 (0.057)         (0.046)
## FDD Lag 7        0.019           0.019
##                 (0.057)         (0.015)
## FDD Lag 8        -0.038         -0.038
##                 (0.057)         (0.034)
## FDD Lag 9        -0.006         -0.006
##                 (0.057)         (0.050)
## FDD Lag 10      -0.112*         -0.112
##                 (0.057)         (0.069)
## FDD Lag 11       -0.063         -0.063
##                 (0.058)         (0.052)
## FDD Lag 12      -0.140**        -0.140*
##                 (0.058)         (0.078)
## Constant        -0.426*         -0.426*
##                 (0.238)         (0.243)
## ----------------------------------------
## Observations      600             600
```

```
## ============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

Let us now use function `estim.armax`, from package `AEC`to estimate an ARMA-X(2,0,1) model:

```
nb.lags <- 1
FDD <- FJ$fdd[(nb.lags+1):T]
names.FDD <- NULL
for(i in 1:nb.lags){
  FDD <- cbind(FDD,FJ$fdd[(nb.lags+1-i):(T-i)])
  names.FDD <- c(names.FDD,paste(" Lag ",toString(i),sep=""))}
colnames(FDD) <- c(" Lag 0",names.FDD)
dprice <- 100*log(price[(k+1):T]/price[1:(T-k)])
dprice <- dprice[(length(dprice)-dim(FDD)[1]+1):length(dprice)]
res.armax <- estim.armax(Y = dprice,p=3,q=0,X=FDD)
```

```
## [1] "================================================="
## [1] "  ESTIMATING"
## [1] "================================================="
## [1] "  END OF ESTIMATION"
## [1] "================================================="
## [1] ""
## [1] "  RESULTS:"
## [1] "  ----------------------"
##                 THETA     st.dev   t.ratio
## c          -0.46556249 0.19554352 -2.380864
## phi   t-1   0.09788977 0.04025907  2.431496
## phi   t-2   0.05049849 0.03827488  1.319364
## phi   t-3   0.07155170 0.03764750  1.900570
## sigma       4.64917949 0.13300769 34.954215
## beta  t-0   0.47015552 0.05665344  8.298800
## beta  t-1   0.10015862 0.05972526  1.676989
## [1] "================================================="
```

Figure 2.6 shows the IRF associated with each of the two models.

```
nb.periods <- 20
IRF1 <- sim.arma(c=0,phi=c(0),theta=eq$coefficients[2:13],sigma=1,
                 T=nb.periods,y.0=c(0),nb.sim=1,make.IRF=1)
IRF2 <- sim.arma(c=0,phi=res.armax$phi,theta=res.armax$beta,sigma=1,
                 T=nb.periods,y.0=rep(0,length(res.armax$phi)),
                 nb.sim=1,make.IRF=1)
par(plt=c(.15,.95,.2,.95))
plot(IRF1,type="l",lwd=2,col="red",xlab="months after shock",
     ylab="Chge in price (percent)")
lines(IRF2,lwd=2,col="red",lty=2)
abline(h=0,col="grey")
```

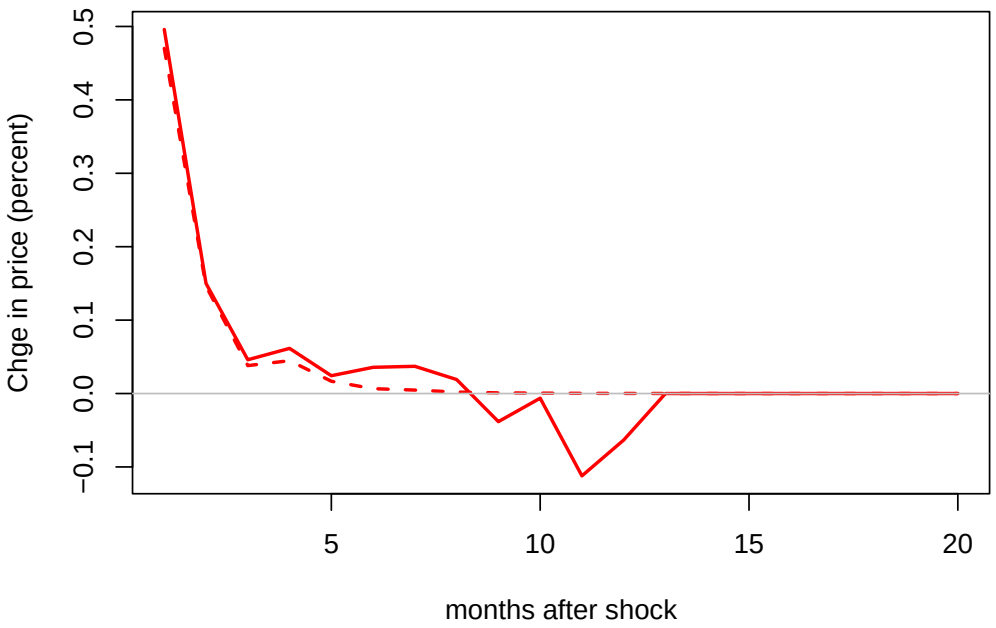


Figure 2.6: Response of changes in orange juice price (in percent) to the number of freezing days. The solid (respectively dashed) line corresponds to the ARMAX(0,0,12) (resp. ARMAX(3,0,1)) model. The first model is estimated by OLS (see above), the second by MLE.

Example 2.2 (Real effect of a monetary policy shock)**.** In this example, we make use of monetary shocks identified through high-frequency data (see

Gertler and Karadi (2015)). This dataset comes from Valerie Ramey's website (see Ramey (2016)).

```
library(AEC)
T <- dim(Ramey)[1]
# Construct growth series:
Ramey$growth <- Ramey$LIP - c(rep(NaN,12),Ramey$LIP[1:(length(Ramey$LIP)-12)])
# Prepare matrix of exogenous variables:
vec.lags <- c(9,12,18)
Matrix.of.Exog <- NULL
shocks <- Ramey$ED2_TC
for(i in 1:length(vec.lags)){Matrix.of.Exog <-
  cbind(Matrix.of.Exog,c(rep(NaN,vec.lags[i]),shocks[1:(T-vec.lags[i])]))}
# Look for dates where data are available:
indic.good.dates <- complete.cases(Matrix.of.Exog)
```

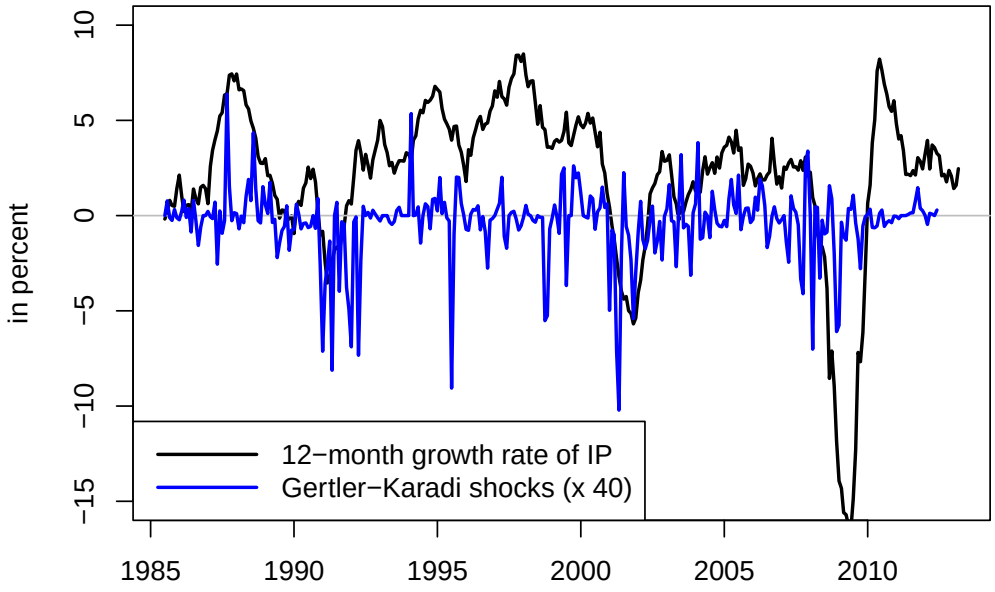


Figure 2.7: The blue line corresponds to monetary-policy shocks identified by means of the Gertler and Karadi (2015)'s approach (high-frequency change in Euro-dollar futures). The black slid line is the year-on-year growth rate of industrial production.

```
# Estimate ARMAX:
p <- 1; q <- 0
x <- estim.armax(Ramey$growth[indic.good.dates],p,q,
                 X=Matrix.of.Exog[indic.good.dates,])
```

```
## [1] "================================================"
## [1] "  ESTIMATING"
## [1] "================================================"
## [1] "  END OF ESTIMATION"
## [1] "================================================"
## [1] ""
## [1] "  RESULTS:"
## [1] "  -----------------------"
##                    THETA       st.dev    t.ratio
## c          -0.0001716198 0.0005845907 -0.2935726
## phi   t-1   0.9825608412 0.0120458531 81.5683897
## sigma       0.0087948724 0.0003211748 27.3834438
## beta  t-0  -0.0193570616 0.0087331529 -2.2165032
## beta  t-1  -0.0225707935 0.0086750938 -2.6017925
## beta  t-2  -0.0070131593 0.0086387440 -0.8118263
## [1] "================================================"
```

```
# Compute IRF:
irf <- sim.arma(0,x$phi,x$beta,x$sigma,T=60,y.0=rep(0,length(x$phi)),
                nb.sim=1,make.IRF=1,X=NaN,beta=NaN)
```

Figure 2.8 displays the resulting IRF, with a 95% confidence band. The code used to produce the confidence bands (i.e., to compute the standard deviation of the dynamic multipliers for the different horizons) is based on the Delta method (see Eq. (**??**)). The codes are available in Appendix 8.6.2.

2.0.8 Maximum Likelihood Estimation of ARMA processes

Consider the general case (of any time series); assume we observe a sample $\mathbf{y} = [y_1, ..., y_T]'$. In order to implement ML techniques (see Section **??**), we

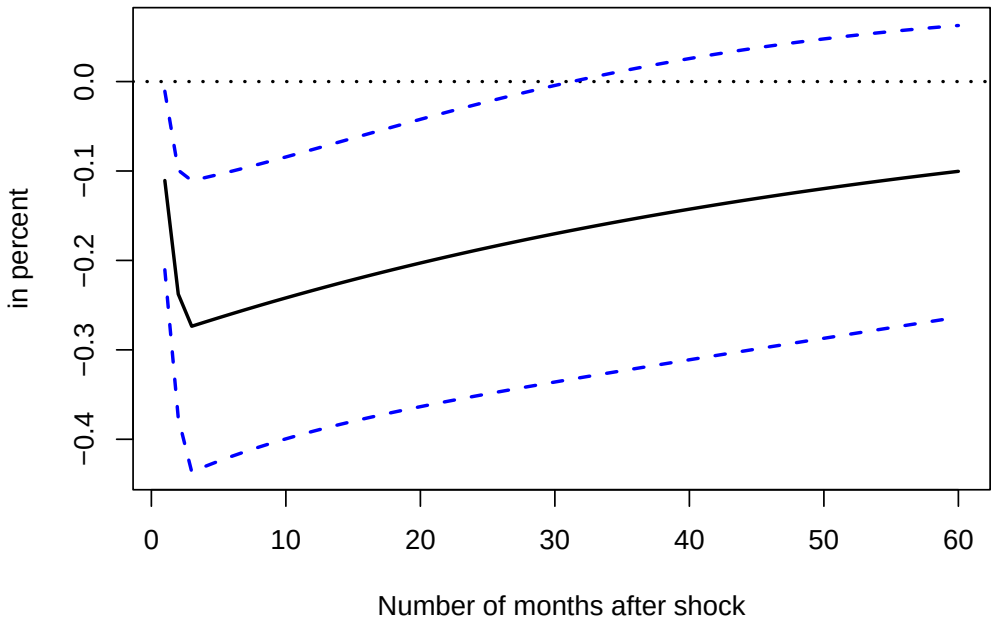


Figure 2.8: Response of industrial-production growth to monetary-policy shocks. Dashed lines correpsond to the ± 2-standard-deviation bands.

need to evaluate the joint p.d.f. (or "likelihood") of $\mathbf{y}$, i.e., $\mathcal{L}(\theta; \mathbf{y})$, where θ is a vector of parameters that characterizes the dynamics of y_t. The Maximum Likelihood (ML) estimate of θ is then given by:

$$\boxed{\theta_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta; \mathbf{y}) = \arg\max_{\theta} \log \mathcal{L}(\theta; \mathbf{y}).}$$

In the time series context, if process y_t is Markovian, then there exists a useful way to rewrite the likelihood $\mathcal{L}(\theta; \mathbf{y})$. Let us first recall the definition of a Markovian process (see also Def. **??**):

Definition 2.10 (Markovian process)**.** Process y_t is Markovian of order one if $f_{Y_t|Y_{t-1},Y_{t-2},...} = f_{Y_t|Y_{t-1}}$. More generally, it is Markovian of order k if $f_{Y_t|Y_{t-1},Y_{t-2},...} = f_{Y_t|Y_{t-1},...,Y_{t-k}}$.

Now, remember Bayes' formula:

$$\mathbb{P}(X_2 = x, X_1 = y) = \mathbb{P}(X_2 = x | X_1 = y)\mathbb{P}(X_1 = y).$$

Using it leads to the following decomposition of our likelihood function:

$$\begin{aligned} f_{Y_T,\dots,Y_1}(y_T,\dots,y_1;\theta) &= f_{Y_T|Y_{T-1},\dots,Y_1}(y_T,\dots,y_1;\theta) \times \\ & \quad f_{Y_{T-1},\dots,Y_1}(y_{T-1},\dots,y_1;\theta). \end{aligned}$$

Using the previous expression recursively, one obtains:

$$f_{Y_T,\dots,Y_1}(y_T,\dots,y_1;\theta) = f_{Y_1}(y_1;\theta)\prod_{t=2}^{T} f_{Y_t|Y_{t-1},\dots,Y_1}(y_t,\dots,y_1;\theta). \tag{2.15}$$

Let us start with the Gaussian AR(1) process (which is Markovian of order one):

$$y_t = c + \phi_1 y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \ i.i.d.\, \mathcal{N}(0,\sigma^2).$$

For $t > 1$:

$$f_{Y_t|Y_{t-1},\dots,Y_1}(y_t,\dots,y_1;\theta) = f_{Y_t|Y_{t-1}}(y_t,y_{t-1};\theta)$$

and

$$f_{Y_t|Y_{t-1}}(y_t,y_{t-1};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_t - c - \phi_1 y_{t-1})^2}{2\sigma^2}\right).$$

These expressions can be plugged into Eq. (2.15). But what about $f_{Y_1}(y_1;\theta)$? There exist two possibilities:

1. **Case 1**: We use the marginal distribution: $y_1 \sim \mathcal{N}\left(\frac{c}{1-\phi_1}, \frac{\sigma^2}{1-\phi_1^2}\right)$.
2. **Case 2**: y_1 is considered to be deterministic. In a way, that means that the first observation is "sacrificed".

For a Gaussian AR(1) process, we have:

1. **Case 1**: The (exact) log-likelihood is:

$$\begin{aligned} \log \mathcal{L}(\theta;\mathbf{y}) &= -\frac{T}{2}\log(2\pi) - T\log(\sigma) + \frac{1}{2}\log(1-\phi_1^2) \\ & \quad -\frac{(y_1 - c/(1-\phi_1))^2}{2\sigma^2/(1-\phi_1^2)} - \sum_{t=2}^{T}\left[\frac{(y_t - c - \phi_1 y_{t-1})^2}{2\sigma^2}\right] \end{aligned} \tag{2.16}$$

   The Maximum Likelihood Estimator of $\theta = [c, \phi_1, \sigma^2]$ is obtained by numerical optimization.

2. **Case 2**: The (conditional) log-likelihood is:

$$\begin{aligned}\log \mathcal{L}^*(\theta; \mathbf{y}) &= -\frac{T-1}{2}\log(2\pi) - (T-1)\log(\sigma) \\ &\quad -\sum_{t=2}^{T}\left[\frac{(y_t - c - \phi_1 y_{t-1})^2}{2\sigma^2}\right]. \end{aligned} \tag{2.17}$$

Exact MLE and conditional MLE have the same asymptotic (i.e. large-sample) distribution. Indeed, when the process is stationary, $f_{Y_1}(y_1;\theta)$ makes a relatively negligible contribution to $\log \mathcal{L}(\theta; \mathbf{y})$.

The conditional MLE has a substantial advantage: in the Gaussian case, the conditional MLE is simply obtained by OLS. Indeed, let us introduce the notations:

$$Y = \begin{bmatrix} y_2 \\ \vdots \\ y_T \end{bmatrix} \quad and \quad X = \begin{bmatrix} 1 & y_1 \\ \vdots & \vdots \\ 1 & y_{T-1} \end{bmatrix}.$$

Eq. (2.17) then rewrites:

$$\begin{aligned}\log \mathcal{L}^*(\theta; \mathbf{y}) &= -\frac{T-1}{2}\log(2\pi) - (T-1)\log(\sigma) \\ &\quad -\frac{1}{2\sigma^2}(Y - X[c, \phi_1]')'(Y - X[c, \phi_1]'), \end{aligned} \tag{2.18}$$

which is maximised for:

$$[\hat{c}, \hat{\phi}_1]' = (X'X)^{-1}X'Y \tag{2.19}$$

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{T-1}\sum_{t=2}^{T}(y_t - \hat{c} - \hat{\phi}_1 y_{t-1})^2 \\ &= \frac{1}{T-1}Y'(I - X(X'X)^{-1}X')Y. \end{aligned} \tag{2.20}$$

Let us turn to the case of an AR(p) process. We have:

$$\begin{aligned}\log \mathcal{L}(\theta; \mathbf{y}) &= \log f_{Y_p,\dots,Y_1}(y_p, \dots, y_1; \theta) + \\ &\quad \underbrace{\sum_{t=p+1}^{T} \log f_{Y_t|Y_{t-1},\dots,Y_{t-p}}(y_t, \dots, y_{t-p}; \theta)}_{\log \mathcal{L}^*(\theta;\mathbf{y})}. \end{aligned}$$

where $f_{Y_p,\dots,Y_1}(y_p,\dots,y_1;\theta)$ is the marginal distribution of $\mathbf{y}_{1:p} := [y_p,\dots,y_1]'$. The marginal distribution of $\mathbf{y}_{1:p}$ is Gaussian; it is therefore fully characterised by its mean and covariance matrix:

$$\begin{aligned}\mathbb{E}(\mathbf{y}_{1:p}) &= \frac{c}{1-\phi_1-\cdots-\phi_p}\mathbf{1}_{p\times 1}\\ \mathbb{V}ar(\mathbf{y}_{1:p}) &= \begin{bmatrix}\gamma_0 & \gamma_1 & \cdots & \gamma_{p-1}\\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2}\\ \vdots & & \ddots & \vdots\\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0\end{bmatrix},\end{aligned}$$

where the γ_i's are computed using the Yule-Walker equations (Eq. (2.9)). Note that they depend, in a non-linear way, on the model parameters. Hence, the maximization of the exact log-likelihood necessitates numerical oprimization procedures. By contrast, the maximization of the conditional log-likelihood $\log\mathcal{L}^*(\theta;\mathbf{y})$ only requires OLS, using Eqs. (2.19) and (2.20), with:

$$Y=\begin{bmatrix}y_{p+1}\\ \vdots\\ y_T\end{bmatrix} \quad and \quad X=\begin{bmatrix}1 & y_p & \cdots & y_1\\ \vdots & \vdots & & \vdots\\ 1 & y_{T-1} & \cdots & y_{T-p}\end{bmatrix}.$$

Again, for stationary processes, conditional and exact MLE have the same asymptotic (large-sample) distribution. In small samples, the OLS formula is however biased. Indeed, consider the regression (where y_t follows an AR(p) process):

$$y_t=\beta'\mathbf{x}_t+\varepsilon_t, \tag{2.21}$$

with $\mathbf{x}_t=[1,y_{t-1},\dots,y_{t-p}]'$ and $\beta=[c,\phi_1,\dots,\phi_p]'$.

The bias results from the fact that $\mathbf{x}_t$ correlates to the ε_s's for $s<t$. To be sure:

$$\mathbf{b}=\beta+(X'X)^{-1}X'\varepsilon, \tag{2.22}$$

and because of the specific form of X, we have non-zero correlation between $\mathbf{x}_t$ and ε_s for $s<t$, therefore $\mathbb{E}[(X'X)^{-1}X'\varepsilon]\neq 0$. Again, asymptotically, the previous expectation goes to zero, and we have:

Proposition 2.9 (Large-sample porperties of the OLS estimator of AR(p) models)**.** *Assume $\{y_t\}$ follows the AR(p) process:*

$$y_t=c+\phi_1y_{t-1}+\cdots+\phi_py_{t-p}+\varepsilon_t$$

where $\{\varepsilon_t\}$ *is an i.i.d. white noise process. If* $\mathbf{b}$ *is the OLS estimator of* β *(Eq.* (2.21)*), we have:*

$$\sqrt{T}(\mathbf{b}-\beta) = \underbrace{\left[\frac{1}{T}\sum_{t=p}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}}_{\xrightarrow{p}\mathbf{Q}^{-1}} \underbrace{\sqrt{T}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t\right]}_{\xrightarrow{d}\mathcal{N}(0,\sigma^2\mathbf{Q})},$$

where $\mathbf{Q} = plim\ \frac{1}{T}\sum_{t=p}^{T}\mathbf{x}_t\mathbf{x}_t' = plim\ \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ *is given by:*

$$\mathbf{Q} = \begin{bmatrix} 1 & \mu & \mu & ... & \mu \\ \mu & \gamma_0+\mu^2 & \gamma_1+\mu^2 & ... & \gamma_{p-1}+\mu^2 \\ \mu & \gamma_1+\mu^2 & \gamma_0+\mu^2 & ... & \gamma_{p-2}+\mu^2 \\ \vdots & \vdots & \vdots & ... & \vdots \\ \mu & \gamma_{p-1}+\mu^2 & \gamma_{p-2}+\mu^2 & ... & \gamma_0+\mu^2 \end{bmatrix}. \tag{2.23}$$

Proof. Rearranging Eq. (3.12), we have:

$$\sqrt{T}(\mathbf{b}-\beta) = (X'X/T)^{-1}\sqrt{T}(X'\varepsilon/T).$$

Let us consider the autocovariances of $\mathbf{v}_t = \mathbf{x}_t\varepsilon_t$, denoted by γ_j^v. Using the fact that $\mathbf{x}_t$ is a linear combination of past ε_t's and that ε_t is a white noise, we get that $\mathbb{E}(\varepsilon_t\mathbf{x}_t) = 0$. Therefore

$$\gamma_j^v = \mathbb{E}(\varepsilon_t\varepsilon_{t-j}\mathbf{x}_t\mathbf{x}_{t-j}').$$

If $j > 0$, we have

$$\begin{aligned}\mathbb{E}(\varepsilon_t\varepsilon_{t-j}\mathbf{x}_t\mathbf{x}_{t-j}') &= \mathbb{E}(\mathbb{E}[\varepsilon_t\varepsilon_{t-j}\mathbf{x}_t\mathbf{x}_{t-j}'|\varepsilon_{t-j},\mathbf{x}_t,\mathbf{x}_{t-j}]) \\ &= \mathbb{E}(\varepsilon_{t-j}\mathbf{x}_t\mathbf{x}_{t-j}'\mathbb{E}[\varepsilon_t|\varepsilon_{t-j},\mathbf{x}_t,\mathbf{x}_{t-j}]) = 0.\end{aligned}$$

Note that, for $j > 0$, we have $\mathbb{E}[\varepsilon_t|\varepsilon_{t-j},\mathbf{x}_t,\mathbf{x}_{t-j}] = 0$ because $\{\varepsilon_t\}$ is an i.i.d. white noise sequence. If $j = 0$, we have:

$$\gamma_0^v = \mathbb{E}(\varepsilon_t^2\mathbf{x}_t\mathbf{x}_t') = \mathbb{E}(\varepsilon_t^2)\mathbb{E}(\mathbf{x}_t\mathbf{x}_t') = \sigma^2\mathbf{Q}.$$

The convergence in distribution of $\sqrt{T}(X'\varepsilon/T) = \sqrt{T}\frac{1}{T}\sum_{t=1}^{T}v_t$ results from Theorem 1.1 (applied on $\mathbf{v}_t = \mathbf{x}_t\varepsilon_t$), using the γ_j^v computed above. □

These two cases (exact or conditional log-likelihoods) can be implemented when asking R to fit an AR process by means of function `arima`. Let us for instance use the output gap of the `US3var` dataset (US quarterly data, covering the period 1959:2 to 2015:1, used in Gouriéroux et al. (2017)).

```
library(AEC)
y <- US3var$y.gdp.gap
ar3.Case1 <- arima(y,order = c(3,0,0),method="ML")
ar3.Case2 <- arima(y,order = c(3,0,0),method="CSS")
rbind(ar3.Case1$coef,ar3.Case2$coef)
```

```
##           ar1         ar2        ar3  intercept
## [1,] 1.191267 -0.08934705 -0.1781163 -0.9226007
## [2,] 1.192003 -0.08811150 -0.1787662 -1.0341696
```

The two sets of estimated coefficients appear to be very close to each other.

Let us now turn to Moving-Average processes. Start with the MA(1):

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \quad \varepsilon_t \sim i.i.d. \mathcal{N}(0, \sigma^2).$$

The ε_t's are easily computed recursively, starting with $\varepsilon_t = y_t - \mu - \theta_1 \varepsilon_{t-1}$. We obtain:

$$\varepsilon_t = y_t - \theta_1 y_{t-1} + \theta_1^2 y_{t-2}^2 + \cdots + (-1)^{t-1} \theta_1^{t-1} y_1 + (-1)^t \theta_1^t \varepsilon_0.$$

Assume that one wants to recover the sequence of $\{\varepsilon_t\}$'s based on observed values of y_t (from date 1 to date t). One can use the previous expression, but what value should be used for ε_0? If one does not use the true value of ε_0 but 0 (say), one does not obtain ε_t, but only an estimate of it ($\widehat{\varepsilon}_t$, say), with:

$$\widehat{\varepsilon}_t = \varepsilon_t - (-1)^t \theta_1^t \varepsilon_0.$$

Clearly, if $|\theta_1| < 1$, then the error becomes small for large t. Formally, when $|\theta_1| < 1$, we have:

$$\widehat{\varepsilon}_t \overset{p}{\to} \varepsilon_t.$$

Hence, when $|\theta_1| < 1$, a consistent estimate of the conditional log-likelihood is given by:

$$\log \widehat{\mathcal{L}}^*(\theta; \mathbf{y}) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) - \sum_{t=1}^{T} \frac{\widehat{\varepsilon}_t^2}{2\sigma^2}. \tag{2.24}$$

Loosely speaking, if $|\theta_1| < 1$ and if T is sufficiently large:

$$\text{approximate conditional MLE} \approx \text{exact MLE}.$$

Note that $\hat{\mathcal{L}}^*(\theta; \mathbf{y})$ is a complicated nonlinear function of μ and θ. Its maximization therefore has to be based on numerical optimization procedures.

Let us not consider the case of a Gaussian MA(q) process:

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \quad \varepsilon_t \sim i.i.d. \mathcal{N}(0, \sigma^2). \tag{2.25}$$

Let us assume that this process is an **invertible MA process**. That is, assume that the roots of:

$$\lambda^q + \theta_1 \lambda^{q-1} + \cdots + \theta_{q-1}\lambda + \theta_q = 0 \tag{2.26}$$

lie strictly inside of the unit circle. In this case, the polynomial form $\Theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$ is *invertible* and Eq. (2.25) writes:

$$\varepsilon_t = \Theta(L)^{-1}(y_t - \mu),$$

which implies that, if we knew all past values of y_t, we would also know ε_t. In this case, we can consistently estimate the ε_t's by recursively computing the $\hat{\varepsilon}_t$'s as follows (for $t > 0$):

$$\hat{\varepsilon}_t = y_t - \mu - \theta_1 \hat{\varepsilon}_{t-1} - \cdots - \theta_q \hat{\varepsilon}_{t-q}, \tag{2.27}$$

with

$$\hat{\varepsilon}_0 = \cdots = \hat{\varepsilon}_{-q+1} = 0. \tag{2.28}$$

In this context, a consistent estimate of the conditional log-likelihood is still given by Eq. (2.24), using Eqs. (2.27) and (2.28) to recursively compute the $\hat{\varepsilon}_t$'s.

Note that we could determine the exact likelihood of an MA process. Indeed, vector $\mathbf{y} = [y_1, \ldots, y_T]'$ is a Gaussian-distributed vector of mean $\mu = [\mu, \ldots, \mu]'$ and of variance:

$$\Omega = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_q & 0 & \cdots & 0 \\ \gamma_1 & \gamma_0 & \gamma_1 & & \ddots & \ddots & \vdots \\ \vdots & \gamma_1 & \ddots & \ddots & & \ddots & 0 \\ \gamma_q & & \ddots & & & & \gamma_q \\ 0 & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \gamma_1 & \gamma_0 & \gamma_1 \\ 0 & \cdots & 0 & \gamma_q & \cdots & \gamma_1 & \gamma_0 \end{bmatrix},$$

where the γ_j's are given by Eq. (2.1). The p.d.f. of $\mathbf{y}$ is then given by (see Prop. 8.18):

$$(2\pi)^{-T/2}|\Omega|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{y}-\mu)'\Omega^{-1}(\mathbf{y}-\mu)\right).$$

For large samples, the computation of this likelihood however becomes numerically demanding.

Finally, let us consider the MLE of an ARMA(p,q) processes:

$$y_t = c+\phi_1 y_{t-1}+\cdots+\phi_p y_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\cdots+\theta_q\varepsilon_{t-q},\ \varepsilon_t \sim i.i.d.\,\mathcal{N}(0,\sigma^2).$$

If the MA part of this process is invertible, the log-likelihood function can be consistently approximated by its conditional counterpart (of the form of Eq. (2.24)), using consistent estimates $\widehat{\varepsilon}_t$ of the ε_t. The $\widehat{\varepsilon}_t$'s are computed recursively as:

$$\widehat{\varepsilon}_t = y_t - c - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} - \theta_1\widehat{\varepsilon}_{t-1} - \cdots - \theta_q\widehat{\varepsilon}_{t-q}, \tag{2.29}$$

given some initial conditions, for instance:

a. $\widehat{\varepsilon}_0 = \cdots = \widehat{\varepsilon}_{-q+1} = 0$ and $y_0 = \cdots = y_{-p+1} = \mathbb{E}(y_i) = \mu$. (Recursions in Eq. (2.29) then start for $t = 1$.)
b. $\widehat{\varepsilon}_p = \cdots = \widehat{\varepsilon}_{p-q+1} = 0$ and actual values of the y_i's for $i \in [1,p]$. In that case, the first p observations of y_t will not be used. Recursions in Eq. (2.29) then start for $t = p+1$.

2.0.9 Specification choice

The previouss section explains how to fit a given ARMA specification. But how to choose an appropriate specification? A possibility is to employ the (P)ACF approach (see Figure 2.2). However, the previous approach leads to either an AR or a MA process (and not an ARMA process). If one wants to consider various ARMA(p,q) specifications, for $p \in \{1,\dots,P\}$ and $q \in \{1,\dots,Q\}$, say, then one can resort to **information criteria**.

In general, when choosing a specification, one faces the following dilemma:

a. Too rich a specification may lead to "overfitting"/misspecification, implying additional estimation errors (in out-of-sample forecasts).

b. Too simple a specification may lead to potential omission of valuable information (e.g., contained in older lags).

The lag selection approach based on the so-called **information criteria** consists in maximizing the fit of the data, but adding a penalty for the "richness" of the model. More precisely, using this approach amounts to minimizing a loss function that (a) negatively depends on the fitting errors and (b) positively depends on the number of parameters in the model.

Definition 2.11 (Information Criteria)**.** The Akaike (AIC), Hannan-Quinn (HQ) and Schwarz information (BIC) criteria are of the form

$$c^{(i)}(k) = \underbrace{\frac{-2\log \mathcal{L}(\hat{\theta}_T(k); \mathbf{y})}{T}}_{\text{decreases w.r.t. } k} + \underbrace{\frac{k\phi^{(i)}(T)}{T}}_{\text{increases w.r.t. } k},$$

with $(i) \in \{AIC, HQ, BIC\}$ and where $\hat{\theta}_T(k)$ denotes the ML estimate of $\theta_0(k)$, which is a vector of parameters of length k.

Criterion (i)		$\phi^{(i)}(T)$
Akaike	AIC	2
Hannan-Quinn	HQ	$2\log(\log(T))$
Schwarz	BIC	$\log(T)$

The lag suggested by criterion (i) is then given by:

$$\boxed{\hat{k}^{(i)} = \underset{k}{\operatorname{argmin}} \quad c^{(i)}(k).}$$

In the case of an ARMA(p,q) process, $k = 2 + p + q$.

Proposition 2.10 (Consistency of the criteria-based lag selection)**.** *The lag selection procedure is consistent (see Def. 8.8) if*

$$\lim_{T\to\infty} \phi(T) = \infty \quad \textit{and} \quad \lim_{T\to\infty} \phi(T)/T = 0.$$

This is notably the case of the HQ and the BIC criteria.

Proof. The true number of lags is denoted by k_0. We will show that $\lim_{T\to\infty} \mathbb{P}(\hat{k}_T \neq k_0) = 0$.

- Case $k < k_0$: The model with k parameter is misspecified, therefore:

  $$\text{plim}_{T\to\infty} \log \mathcal{L}(\hat{\theta}_T(k); \mathbf{y})/T < \text{plim}_{T\to\infty} \log \mathcal{L}(\hat{\theta}_T(k_0); \mathbf{y})/T.$$

  Hence, if $\lim_{T\to\infty} \phi(T)/T = 0$, we have: $\lim_{T\to\infty} \mathbb{P}(c(k_0) \geq c(k)) \to 0$ and

  $$\lim_{T\to\infty} \mathbb{P}(\hat{k} < k_0) \leq \lim_{T\to\infty} \mathbb{P}\left\{c(k_0) \geq c(k) \text{ for some } k < k_0\right\} = 0.$$

- Case $k > k_0$: under the null hypothesis, the likelihood ratio (LR) test statistic (see Def. **??**) satisfies:

  $$2\left(\log \mathcal{L}(\hat{\theta}_T(k); \mathbf{y}) - \log \mathcal{L}(\hat{\theta}_T(k_0); \mathbf{y})\right) \sim \chi^2(k - k_0).$$

  If $\lim_{T\to\infty} \phi(T) = \infty$, we have: $\text{plim}_{T\to\infty} -2\left(\log \mathcal{L}(\hat{\theta}_T(k); \mathbf{y}) - \log \mathcal{L}(\hat{\theta}_T(k_0); \mathbf{y})\right)/\phi(T) = 0$. Hence $\text{plim}_{T\to\infty} T[c(k_0) - c(k)]/\phi(T) \leq -1$ and $\lim_{T\to\infty} \mathbb{P}(c(k_0) \geq c(k)) \to 0$, which implies, in the same spirit as before, that $\lim_{T\to\infty} \mathbb{P}(\hat{k} > k_0) = 0$.

Therefore, $\lim_{T\to\infty} \mathbb{P}(\hat{k} = k_0) = 1$. □

Example 2.3 (Linear regression)**.** Consider a linear regression with normal disturbances:

$$y_t = \mathbf{x}_t'\beta + \varepsilon_t, \quad \varepsilon_t \sim i.i.d.\mathcal{N}(0, \sigma^2).$$

The associated log-likelihood is of the form of Eq. (2.24). In that case, we have:

$$\begin{aligned} c^{(i)}(k) &= \frac{-2\log \mathcal{L}(\hat{\theta}_T(k); \mathbf{y})}{T} + \frac{k\phi^{(i)}(T)}{T} \\ &\approx \log(2\pi) + \log(\widehat{\sigma^2}) + \frac{1}{T}\sum_{t=1}^{T} \frac{\varepsilon_t^2}{\widehat{\sigma^2}} + \frac{k\phi^{(i)}(T)}{T}. \end{aligned}$$

For a large T, for all consistent estimation scheme, we have:

$$\widehat{\sigma^2} \approx \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t^2 = SSR/T.$$

Hence $\hat{k}^{(i)} \approx \underset{k}{\text{argmin}} \quad \log(SSR/T) + \dfrac{k\phi^{(i)}(T)}{T}$.

Example 2.4 (Swiss GDP growth)**.** Consider a long historical time series of the Swiss GDP growth (see Figure 1.3), taken from the Jordà et al. (2017) dataset. Let us look for the best ARMA specification using the AIC criteria:

```
library(AEC);data(JST)
data <- subset(JST,iso=="CHE")
T <- dim(data)[1]
y <- c(NaN,log(data$gdp[2:T]/data$gdp[1:(T-1)]))
# Use AIC criteria to look for appropriate specif:
max.p <- 3;max.q <- 3;
all.AIC <- NULL
for(p in 0:max.p){
  for(q in 0:max.q){
    res <- arima(y,order=c(p,0,q))
    if(res$aic<min(all.AIC)){best.p<-p;best.q<-q}
    all.AIC <- c(all.AIC,res$aic)}}
print(c(best.p,best.q))
```

```
## [1] 1 0
```

The best specification therefore is an AR(1) model. That is, although an AR(2) (say) would result in a better fit of the data, the fit improvement is not be large enough to compensate for the additional AIC cost associated with an additional parameter.

Chapter 3

Multivariate models

This section presents Vector Auto-Regressive Moving-Average (SVARMA) models. These models are widely used in macroeconomic analysis. While simple and easy to estimate, they make it possible to conveniently capture the dynamics of complex multivariate systems. VAR popularity is notably due to Sims (1980)'s influential work. A nice survey if proposed by Stock and Watson (2016).

In economics, VAR models are often employed in order to identify *structural* shocks, that are independent primitive exogenous forces that drive economic variables (Ramey (2016)). They are often given a specific economic meaning (e.g., demand and supply shocks).

Working with these models (VAR and VARMA models) often involves two steps: in a first step, the **reduced-form** version of the model is estimated; in a second step, **structural shocks** are identified and IRFs are produced.

3.0.1 Definition of VARs (and SVARMA) models

Definition 3.1 ((S)VAR model)**.** Let y_t denote a $n \times 1$ vector of random variables. Process y_t follows a p^{th}-order (S)VAR if, for all t, we have

$$\begin{array}{rrcl} VAR: & y_t & = & c + \Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p} + \varepsilon_t, \\ SVAR: & y_t & = & c + \Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p} + B\eta_t, \end{array} \tag{3.1}$$

with $\varepsilon_t = B\eta_t$, where $\{\eta_t\}$ is a white noise sequence whose components are mutually and serially independent.

The first line of Eq. (3.1) corresponds to the **reduced-form** of the VAR model (**structural form** for the second line).

While the structural shocks (the components of η_t) are mutually uncorrelated, this is not the case of the *innovations*, that are the components of ε_t. However, in boths cases, vectors η_t and ε_t are serially correlated (through time).

As was the case for univariate models, VARs can be extended with MA terms in η_t:

Definition 3.2 ((S)VARMA model)**.** Let y_t denote a $n \times 1$ vector of random variables. Process y_t follows a VARMA model of order (p,q) if, for all t, we have

$$\begin{array}{rrcl} VARMA: & y_t & = & c + \Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p} + \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \cdots + \Theta_q, \\ SVARMA: & y_t & = & c + \Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p} + B_0 \eta_t + B_1 \eta_{t-1} + \cdots + B_q \eta_{t-q}, \end{array} \tag{3.2}$$

with $\varepsilon_t = B_0 \eta_t$ (and $B_j = \Theta_j B_0$, for $j \geq 0$), where $\{\eta_t\}$ is a white noise sequence whose components are mutually and serially independent.

3.0.2 IRFs in SVARMA

One of the main objectives of macro-econometrics is to derive IRFs, that represent the dynamic effects of structural shocks (components of η_t) though the system of variables y_t.

Formally, an IRF is a difference in conditional expectations:

$$\boxed{\Psi_{i,j,h} = \mathbb{E}(y_{i,t+h} | \eta_{j,t} = 1) - \mathbb{E}(y_{i,t+h})}$$

(effect on $y_{i,t+h}$ of a one-unit shock on $\eta_{j,t}$).

If the dynamics of process y_t can be described as a VARMA model, and if y_t is covariance stationary (see Def. 1.4), then y_t admits the following infinite MA representation (MA(∞)):

$$y_t = \mu + \sum_{h=0}^{\infty} \Psi_h \eta_{t-h}. \tag{3.3}$$

This is also the Wold decomposition of process $\{y_t\}$ (see Theorem 2.2).

Estimating IRFs amounts to estimating the Ψ_h's. In general, there exist three main approaches for that:

- Calibrate and solve a (purely structural) Dynamic Stochastic General Equilibrium (DSGE) model at the first order (linearization). The solution takes the form of Eq. (3.3).
- Directly estimate the Ψ_h based on **projection approaches** (see Section **??**).
- Approximate the infinite MA representation by estimating a parsimonious type of model, e.g. **VAR(MA) models** (see Section 3.0.4). Once a (Structural) VARMA representation is obtained, Eq. (3.3) is easily deduced. For that, one can use the same recursive algorithm as for univariate processes (see Prop. 2.7).

Typically, consider the AR(2) case. The first steps of the algorithm mentioned in the last bullet point are as follows:

$$\begin{aligned} y_t &= \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + B\eta_t \\ &= \Phi_1(\Phi_1 y_{t-2} + \Phi_2 y_{t-3} + B\eta_{t-1}) + \Phi_2 y_{t-2} + B\eta_t \\ &= B\eta_t + \Phi_1 B\eta_{t-1} + (\Phi_2 + \Phi_1^2) y_{t-2} + \Phi_1\Phi_2 y_{t-3} \\ &= B\eta_t + \Phi_1 B\eta_{t-1} + (\Phi_2 + \Phi_1^2)(\Phi_1 y_{t-3} + \Phi_2 y_{t-4} + B\eta_{t-2}) + \Phi_1\Phi_2 y_{t-3} \\ &= \underbrace{B}_{=\Psi_0}\eta_t + \underbrace{\Phi_1 B}_{=\Psi_1}\eta_{t-1} + \underbrace{(\Phi_2 + \Phi_1^2)B}_{=\Psi_2}\eta_{t-2} + f(y_{t-3}, y_{t-4}). \end{aligned}$$

In particular, we have $B = \Psi_0$. Matrix B indeed captures the contemporaneous impact of η_t on y_t. That is why matrix B is sometimes called *impulse matrix*.

Example 3.1 (IRFs of an SVARMA model)**.** Consider the following VARMA(1,1) model:

$$y_t = \underbrace{\begin{bmatrix} 0.5 & 0.3 \\ -0.4 & 0.7 \end{bmatrix}}_{\Phi_1} y_{t-1} + \underbrace{\begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}}_{B} \eta_t + \underbrace{\begin{bmatrix} 2 & 0 \\ 1 & 0.5 \end{bmatrix}}_{\Theta_1} \underbrace{\begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}}_{B} \eta_{t-1} \quad (3.4)$$

We can use function `simul.VARMA` of package `AEC` to produce IRFs (using `indic.IRF=1` in the list of arguments):

```
library(AEC)
distri <- list(type=c("gaussian","gaussian"),df=c(4,4))
n <- length(distri$type) # dimension of y_t
nb.sim <- 30
eps <- simul.distri(distri,nb.sim)
Phi <- array(NaN,c(n,n,1))
Phi[,,1] <- matrix(c(.5,-.4,.3,.7),2,2)
p <- dim(Phi)[3]
Theta <- array(NaN,c(n,n,1))
Theta[,,1] <- -matrix(c(2,1,0,.5),2,2)
q <- dim(Theta)[3]
Mu <- rep(0,n)
C <- matrix(c(1,-1,2,1),2,2)
Model <- list(
  Mu = Mu,Phi = Phi,Theta = Theta,C = C,distri = distri)
Y0 <- rep(0,n)
eta0 <- c(1,0)
res.sim.1 <- simul.VARMA(Model,nb.sim,Y0,eta0,indic.IRF=1)
eta0 <- c(0,1)
res.sim.2 <- simul.VARMA(Model,nb.sim,Y0,eta0,indic.IRF=1)
par(plt=c(.15,.95,.25,.8))
par(mfrow=c(2,2))
plot(res.sim.1$Y[1,],las=1,
     type="l",lwd=3,xlab="",ylab="",
     main=expression(paste("Response of ",y[1,"*,*",t],
                           " to a one-unit increase in ",eta[1],sep="")))
abline(h=0,col="grey",lty=3)
plot(res.sim.2$Y[1,],las=1,
     type="l",lwd=3,xlab="",ylab="",
     main=expression(paste("Response of ",y[1,"*,*",t],
                           " to a one-unit increase in ",eta[2],sep="")))
abline(h=0,col="grey",lty=3)
plot(res.sim.1$Y[2,],las=1,
     type="l",lwd=3,xlab="",ylab="",
     main=expression(paste("Response of ",y[2,"*,*",t],
                           " to a one-unit increase in ",eta[1],sep="")))
abline(h=0,col="grey",lty=3)
```

```
plot(res.sim.2$Y[2,],las=1,
     type="l",lwd=3,xlab="",ylab="",
     main=expression(paste("Response of ",y[2,"*,*",t],
                           " to a one-unit increase in ",eta[2],sep="")))
abline(h=0,col="grey",lty=3)
```

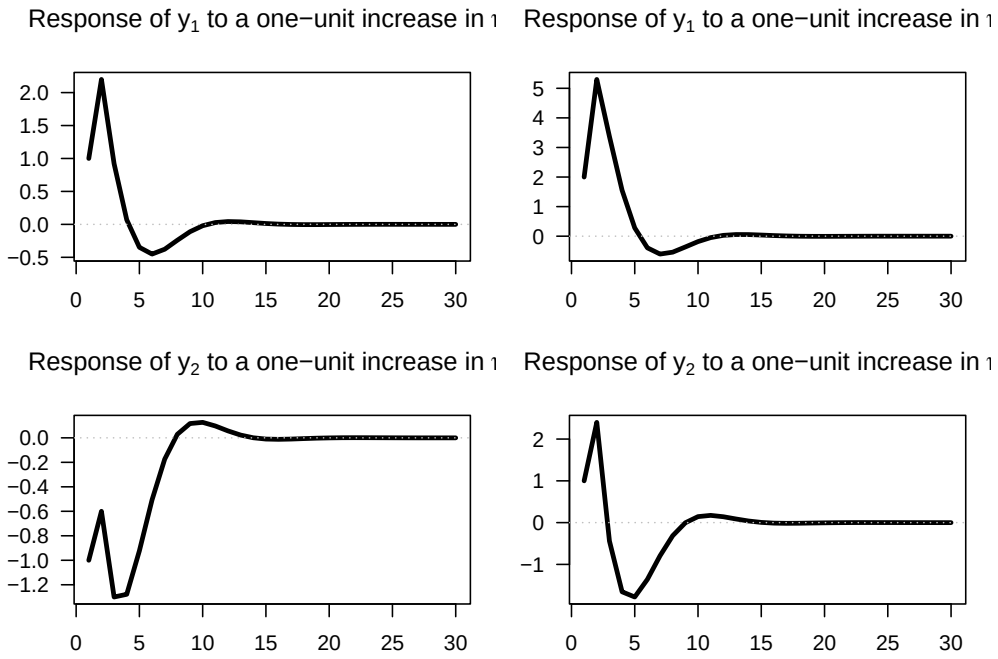


Figure 3.1: Impulse response functions

3.0.3 Covariance-stationary VARMA models

Let's come back to the infinite MA case (Eq. (3.3)):

$$y_t = \mu + \sum_{h=0}^{\infty} \Psi_h \eta_{t-h}.$$

For y_t to be covariance-stationary (and ergodic for the mean), it has to be the case that

$$\sum_{i=0}^{\infty} \|\Psi_i\| < \infty, \tag{3.5}$$

where $\|A\|$ denotes a norm of the matrix A (e.g. $\|A\| = \sqrt{tr(AA')}$). This notably implies that if y_t is stationary (and ergodic for the mean), then $\|\Psi_h\| \to 0$ when h gets large.

What should be satisfied by Φ_k's and Θ_k's for a VARMA-based process (Eq. (??)) to be stationary? The conditions will be similar to that we had in the univariate case (see Prop. 2.5). Let us introduce the following notations:

$$\begin{aligned} y_t &= c + \underbrace{\Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p}}_{\text{AR component}} + \\ &\quad \underbrace{B\eta_t + \Theta_1 B\eta_{t-1} + \cdots + \Theta_q B\eta_{t-q}}_{\text{MA component}} \\ &\Leftrightarrow \underbrace{(I - \Phi_1 L - \cdots - \Phi_p L^p)}_{=\Phi(L)} y_t = c + \underbrace{(I - \Theta_1 L - ... - \Theta_q L^q)}_{=\Theta(L)} B\eta_t. \end{aligned} \tag{3.6}$$

Process y_t is stationary iff the roots of $\det(\Phi(z)) = 0$ are strictly outside the unit circle or, equivalently, iff the eigenvalues of

$$\Phi = \begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_p \\ I & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & I & 0 \end{bmatrix} \tag{3.7}$$

lie strictly within the unit circle. Hence, as was the case for univariate processes, the covariance-stationarity of a VARMA model depends only on the specification of its AR part.

Let's derive the first two unconditional moments of a (covariance-stationary) VARMA process.

Based on Eq. (3.6), we have $\mathbb{E}(\Phi(L)y_t) = c$, which gives $\Phi(1)\mathbb{E}(y_t) = c$, or::

$$\mathbb{E}(y_t) = (I - \Phi_1 - \cdots - \Phi_p)^{-1} c.$$

The autocovariances of y_t can be deduced from the infinite MA representation (Eq. (3.3)). We have:

$$\gamma_j \equiv \mathbb{C}ov(y_t, y_{t-j}) = \sum_{i=j}^{\infty} \Psi_i \Psi'_{i-j}.$$

(Note that this infinite sum exists as soon as Eq. (3.5) is satisfied.)

Conditional means and autocovariances can also be deduced from Eq. (3.3). For $0 \leq h$ and $0 \leq h_1 \leq h_2$:

$$\begin{aligned} \mathbb{E}_t(y_{t+h}) &= \mu + \sum_{k=0}^{\infty} \Psi_{k+h}\eta_{t-k} \\ \mathbb{C}ov_t(y_{t+1+h_1}, y_{t+1+h_2}) &= \sum_{k=0}^{h_1} \Psi_k \Psi'_{k+h_2-h_1}. \end{aligned}$$

The previous formula implies in particular that the forecasting error $y_{t+h} - \mathbb{E}_t(y_{t+h})$ has a variance equal to:

$$\mathbb{V}ar_t(y_{t+h}) = \sum_{k=1}^{h} \Psi_k \Psi'_k.$$

Because the η_t are mutually and serially independent (and therefore uncorrelated), we have:

$$\mathbb{V}ar(\Psi_k \eta_{t-k}) = \mathbb{V}ar\left(\sum_{i=1}^{n} \psi_{k,i}\eta_{i,t-k}\right) = \sum_{i=1}^{n} \psi_{k,i}\psi'_{k,i},$$

where $\psi_{k,i}$ denotes the i^{th} column of Ψ_k.

This suggests the following decomposition of the variance of the forecast error (called **variance decomposition**):

$$\mathbb{V}ar_t(y_{t+h}) = \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{h} \psi_{k,i}\psi'_{k,i}}_{\text{Contribution of } \eta_{i,t}}.$$

Let us now turn to the estimation of VAR(MA) models.

If there is a MA component, OLS regressions yield biased estimates (even for asymptotically large samples).

Assume y_t follows a VARMA(1,1) model. We have:

$$y_{i,t} = \phi_i y_{t-1} + \varepsilon_{i,t},$$

where ϕ_i is the i^{th} row of Φ_1, and where $\varepsilon_{i,t}$ is a linear combination of η_t and η_{t-1}.

Since y_{t-1} (the regressor) is correlated to η_{t-1}, it is also correlated to $\varepsilon_{i,t}$.

The OLS regression of $y_{i,t}$ on y_{t-1} yields a biased estimator of ϕ_i. Hence, SVARMA models cannot be consistently estimated by simple OLS regressions (contrary to VAR models, as we will see in the next section); instrumental-variable approaches can be employed to estimate SVARMA models.

3.0.4 VAR estimation

This section discusses the estimation of VAR models. (The estimation of SVARMA models is more challenging, see, e.g., Gouriéroux et al. (2020).) Eq. (3.1) can be written:

$$y_t = c + \Phi(L)y_{t-1} + \varepsilon_t,$$

with $\Phi(L) = \Phi_1 + \Phi_2 L + \cdots + \Phi_p L^{p-1}$.

Consequently:

$$y_t \mid y_{t-1}, y_{t-2}, \dots, y_{-p+1} \sim \mathcal{N}(c + \Phi_1 y_{t-1} + \dots \Phi_p y_{t-p}, \Omega).$$

Using Hamilton (1994)'s notations, denote with Π the matrix $\begin{bmatrix} c & \Phi_1 & \Phi_2 & \dots & \Phi_p \end{bmatrix}'$ and with x_t the vector $\begin{bmatrix} 1 & y'_{t-1} & y'_{t-2} & \dots & y'_{t-p} \end{bmatrix}'$, we have:

$$y_t = \Pi' x_t + \varepsilon_t. \tag{3.8}$$

The previous representation is convenient to discuss the estimation of the VAR model, as parameters are gathered in two matrices only: Π and Ω.

Let us start with the case where the shocks are Gaussian.

Proposition 3.1 (MLE of a Gaussian VAR)**.** *If y_t follows a VAR(p) (see Definition 3.1), and if $\varepsilon_t \sim$ i.i.d. $\mathcal{N}(0, \Omega)$, then the ML estimate of Π, denoted by $\widehat{\Pi}$ (see Eq. (3.8)), is given by*

$$\widehat{\Pi} = \left[\sum_{t=1}^{T} x_t x'_t\right]^{-1} \left[\sum_{t=1}^{T} y'_t x_t\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{3.9}$$

where $\mathbf{X}$ *is the* $T \times (np)$ *matrix whose* t^{th} *row is* x_t *and where* $\mathbf{y}$ *is the* $T \times n$ *matrix whose* t^{th} *row is* y_t'.

That is, the i^{th} *column of* $\hat{\Pi}$ *(*b_i*, say) is the OLS estimate of* β_i*, where:*

$$y_{i,t} = \beta_i' x_t + \varepsilon_{i,t}, \tag{3.10}$$

(i.e., $\beta_i' = [c_i, \phi_{i,1}', ..., \phi_{i,p}']'$*).*

The ML estimate of Ω*, denoted by* $\widehat{\Omega}$*, coincides with the sample covariance matrix of the* n *series of the OLS residuals in Eq.* (3.10)*, i.e.:*

$$\hat{\Omega} = \frac{1}{T} \sum_{i=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t', \quad \text{with } \hat{\varepsilon}_t = y_t - \hat{\Pi}' x_t. \tag{3.11}$$

The asymptotic distributions of these estimators are the ones resulting from standard OLS formula.

Proof. See Appendix 8.5. □

As stated by Proposition 3.2, when the shocks are not Gaussian, then the OLS regressions still provide consistent estimates of the model parameters. However, since x_t correlates to ε_s for $s < t$, the OLS estimator $\mathbf{b}_i$ of β_i is biased in small sample. (That is also the case for the ML estimator.)

Indeed, denoting by ε_i the $T \times 1$ vector of $\varepsilon_{i,t}$'s, and using the notations of b_i and β_i introduced in Proposition 3.1, we have:

$$\mathbf{b}_i = \beta_i + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon_i. \tag{3.12}$$

We have non-zero correlation between x_t and $\varepsilon_{i,s}$ for $s < t$ and, therefore, $\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon_i] \neq 0$.

However, when y_t is covariance stationary, then $\frac{1}{n}\mathbf{X}'\mathbf{X}$ converges to a positive definite matrix $\mathbf{Q}$, and $\frac{1}{n}X'\varepsilon_i$ converges to 0. Hence $\mathbf{b}_i \xrightarrow{p} \beta_i$. More precisely:

Proposition 3.2 (Asymptotic distribution of the OLS estimate of β_i)**.** *If* y_t *follows a VAR model, as defined in Definition 3.1, we have:*

$$\sqrt{T}(\mathbf{b}_i - \beta_i) = \underbrace{\left[\frac{1}{T}\sum_{t=p}^{T} x_t x_t'\right]^{-1}}_{\xrightarrow{p} \mathbf{Q}^{-1}} \underbrace{\sqrt{T}\left[\frac{1}{T}\sum_{t=1}^{T} x_t \varepsilon_{i,t}\right]}_{\xrightarrow{d} \mathcal{N}(0, \sigma_i^2 \mathbf{Q})},$$

where $\sigma_i = \mathbb{V}ar(\varepsilon_{i,t})$ *and where* $\mathbf{Q} = plim\ \frac{1}{T}\sum_{t=p}^{T} x_t x_t'$ *is given by:*

$$\mathbf{Q} = \begin{bmatrix} 1 & \mu' & \mu' & ... & \mu' \\ \mu & \gamma_0 + \mu\mu' & \gamma_1 + \mu\mu' & ... & \gamma_{p-1} + \mu\mu' \\ \mu & \gamma_1 + \mu\mu' & \gamma_0 + \mu\mu' & ... & \gamma_{p-2} + \mu\mu' \\ \vdots & \vdots & \vdots & ... & \vdots \\ \mu & \gamma_{p-1} + \mu\mu' & \gamma_{p-2} + \mu\mu' & ... & \gamma_0 + \mu\mu' \end{bmatrix}. \tag{3.13}$$

Proof. See Appendix 8.5. □

The following proposition extends the previous proposition and includes covariances between different β_i's as well as the asymptotic distribution of the ML estimates of Ω.

Proposition 3.3 (Asymptotic distribution of the OLS estimates)**.** *If* y_t *follows a VAR model, as defined in Definition 3.1, we have:*

$$\sqrt{T}\begin{bmatrix} vec(\hat{\Pi} - \Pi) \\ vec(\hat{\Omega} - \Omega) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Omega \otimes \mathbf{Q}^{-1} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}\right), \tag{3.14}$$

where the component of Σ_{22} *corresponding to the covariance between* $\hat{\sigma}_{i,j}$ *and* $\hat{\sigma}_{k,l}$ *(for* $i, j, l, m \in \{1, ..., n\}^4$*) is equal to* $\sigma_{i,l}\sigma_{j,m} + \sigma_{i,m}\sigma_{j,l}$.

Proof. See Hamilton (1994), Appendix of Chapter 11. □

Naturally, in practice, Ω is replaced with $\hat{\Omega}$, $\mathbf{Q}$ is replaced with $\hat{\mathbf{Q}} = \frac{1}{T}\sum_{t=p}^{T} x_t x_t'$ and Σ with the matrix whose components are of the form $\hat{\sigma}_{i,l}\hat{\sigma}_{j,m} + \hat{\sigma}_{i,m}\hat{\sigma}_{j,l}$, where the $\hat{\sigma}_{i,l}$'s are the components of $\hat{\Omega}$.

The simplicity of the VAR framework and the tractability of its MLE open the way to convenient econometric testing. Let's illustrate this with the likelihood ratio test (see Def. **??**). The maximum value achieved by the MLE is

$$\log \mathcal{L}(Y_T; \hat{\Pi}, \hat{\Omega}) = -\frac{Tn}{2}\log(2\pi) + \frac{T}{2}\log\left|\hat{\Omega}^{-1}\right| - \frac{1}{2}\sum_{t=1}^{T}\left[\hat{\varepsilon}_t'\hat{\Omega}^{-1}\hat{\varepsilon}_t\right].$$

The last term is:

$$\begin{aligned}\sum_{t=1}^{T}\hat{\varepsilon}_t'\hat{\Omega}^{-1}\hat{\varepsilon}_t &= \text{Tr}\left[\sum_{t=1}^{T}\hat{\varepsilon}_t'\hat{\Omega}^{-1}\hat{\varepsilon}_t\right] = \text{Tr}\left[\sum_{t=1}^{T}\hat{\Omega}^{-1}\hat{\varepsilon}_t\hat{\varepsilon}_t'\right] \\ &= \text{Tr}\left[\hat{\Omega}^{-1}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t'\right] = \text{Tr}\left[\hat{\Omega}^{-1}\left(T\hat{\Omega}\right)\right] = Tn.\end{aligned}$$

Therefore, the optimized log-likelihood is simply obtained by:

$$\log\mathcal{L}(Y_T;\hat{\Pi},\hat{\Omega}) = -(Tn/2)\log(2\pi) + (T/2)\log\left|\hat{\Omega}^{-1}\right| - Tn/2. \tag{3.15}$$

Assume that we want to test the null hypothesis that a set of variables follows a VAR(p_0) against the alternative specification of p_1 ($> p_0$).

Let us denote by $\hat{L}_0$ and $\hat{L}_1$ the maximum log-likelihoods obtained with p_0 and p_1 lags, respectively.

Under the null hypothesis (H_0: $p = p_0$), we have:

$$2\left(\hat{L}_1 - \hat{L}_0\right) = T\left(\log\left|\hat{\Omega}_1^{-1}\right| - \log\left|\hat{\Omega}_0^{-1}\right|\right) \sim \chi^2(n^2(p_1 - p_0)).$$

What precedes can be used to help determine the appropriate number of lags to use in the specification. In a VAR, using too many lags consumes numerous degrees of freedom: with p lags, each of the n equations in the VAR contains $n \times p$ coefficients plus the intercept term. Adding lags improve in-sample fit, but is likely to result in over-parameterization and affect the **out-of-sample** prediction performance.

To select appropriate lag length, **selection criteria** can be used (see Definition 2.11). In the context of VAR models, using Eq. (3.15), we have:

$$\begin{aligned}AIC &= cst + \log\left|\hat{\Omega}\right| + \frac{2}{T}N \\ BIC &= cst + \log\left|\hat{\Omega}\right| + \frac{\log T}{T}N,\end{aligned}$$

where $N = p \times n^2$.

3.0.5 Block exogeneity and Granger causality

Block exogeneity

Let's decompose y_t into two subvectors $y_t^{(1)}$ $(n_1 \times 1)$ and $y_t^{(2)}$ $(n_2 \times 1)$, with $y_t' = [y_t^{(1)'}, y_t^{(2)'}]$ (and therefore $n = n_1 + n_2$), such that:

$$\begin{bmatrix} y_t^{(1)} \\ y_t^{(2)} \end{bmatrix} = \begin{bmatrix} \Phi^{(1,1)} & \Phi^{(1,2)} \\ \Phi^{(2,1)} & \Phi^{(2,2)} \end{bmatrix} \begin{bmatrix} y_{t-1}^{(1)} \\ y_{t-1}^{(2)} \end{bmatrix} + \varepsilon_t.$$

Using, e.g., a likelihood ratio test (see Def. **??**), one can easily test for block exogeneity of $y_t^{(2)}$ (say). The null assumption can be expressed as $\Phi^{(2,1)} = 0$.

Granger Causality

Granger (1969) developed a method to explore **causal relationships** among variables. The approach consists in determining whether the past values of $y_{1,t}$ can help explain the current $y_{2,t}$ (beyond the information already included in the past values of $y_{2,t}$).

Formally, let us denote three information sets:

$$\begin{aligned} \mathcal{J}_{1,t} &= \{y_{1,t}, y_{1,t-1}, ...\} \\ \mathcal{J}_{2,t} &= \{y_{2,t}, y_{2,t-1}, ...\} \\ \mathcal{J}_{t} &= \{y_{1,t}, y_{1,t-1}, ... y_{2,t}, y_{2,t-1}, ...\}. \end{aligned}$$

We say that $y_{1,t}$ Granger-causes $y_{2,t}$ if

$$\mathbb{E}\left[y_{2,t} \mid \mathcal{J}_{2,t-1}\right] \neq \mathbb{E}\left[y_{2,t} \mid \mathcal{J}_{t-1}\right].$$

To get the intuition behind the testing procedure, consider the following bivariate VAR(p) process:

$$\begin{aligned} y_{1,t} &= c_1 + \Sigma_{i=1}^{p} \Phi_i^{(11)} y_{1,t-i} + \Sigma_{i=1}^{p} \Phi_i^{(12)} y_{2,t-i} + \varepsilon_{1,t} \\ y_{2,t} &= c_2 + \Sigma_{i=1}^{p} \Phi_i^{(21)} y_{1,t-i} + \Sigma_{i=1}^{p} \Phi_i^{(22)} y_{2,t-i} + \varepsilon_{2,t}, \end{aligned}$$

where $\Phi_k^{(ij)}$ denotes the element (i, j) of Φ_k.

Then, $y_{1,t}$ is said not to Granger-cause $y_{2,t}$ if

$$\Phi_1^{(21)} = \Phi_2^{(21)} = ... = \Phi_p^{(21)} = 0.$$

Therefore the hypothesis testing is

$$\begin{cases} H_0: & \Phi_1^{(21)} = \Phi_2^{(21)} = ... = \Phi_p^{(21)} = 0 \\ H_1: & \Phi_1^{(21)} \neq 0 \text{ or } \Phi_2^{(21)} \neq 0 \text{ or } ... \Phi_p^{(21)} \neq 0. \end{cases}$$

Loosely speaking, we reject H_0 if some of the coefficients on the lagged $y_{1,t}$'s are statistically significant. Formally, this can be tested using the F-test or asymptotic chi-square test. The F-statistic is

$$F = \frac{(RSS - USS)/p}{USS/(T - 2p - 1)},$$

where RSS is the Restricted sum of squared residuals and USS is the Unrestricted sum of squared residuals. Under H_0, the F-statistic is distributed as $\mathcal{F}(p, T - 2p - 1)$. (We have $pF \underset{T\to\infty}{\to} \chi^2(p)$.)

3.0.6 Identification problem and standard identification techniques

In Section 3.0.4, we have seen how to estimate $\mathbb{V}ar(\varepsilon_t) = \Omega$ and the Φ_k matrices in the context of a VAR model. But the IRFs are functions of B and the Φ_k's, not of Ω the Φ_k's (see Section 3.0.2). We have $\Omega = BB'$, but this is not sufficient to recover B.

Indeed, seen a system of equations whose unknowns are the $b_{i,j}$'s (components of B), the system $\Omega = BB'$ contains only $n(n+1)/2$ linearly independent equations. For instance, for $n = 2$:

$$\begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{21} \\ b_{12} & b_{22} \end{bmatrix}$$
$$\Leftrightarrow \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{bmatrix} = \begin{bmatrix} b_{11}^2 + b_{12}^2 & b_{11}b_{21} + b_{12}b_{22} \\ b_{11}b_{21} + b_{12}b_{22} & b_{22}^2 + b_{21}^2 \end{bmatrix}.$$

We then have 3 linearly independent equations but 4 unknowns. Therefore, B is not identified based on second-order moments. Additional restrictions are required to identify B. This section covers two standard identification schemes: **short-run** and **long-run** restrictions:

1. A **short-run restriction (SRR)** prevents a structural shock from affecting an endogenous variable contemporaneously.

- Easy to implement: the appropriate entries of B are set to 0.
- Particular case: **Cholesky, or recursive approach**.
- Examples: Bernanke (1986), Sims (1986), Galí (1992), Ruibio-Ramírez et al. (2010).

2. A **long-run restriction (LRR)** prevents a structural shock from having a cumulative impact on one of the endogenous variables.

- Additional computations are required to implement this. One needs to compute the cumulative effect of one of the structural shocks u_t on one of the endogenous variable.
- Examples: Blanchard and Quah (1989), Faust and Leeper (1997), Galí (1999), Erceg et al. (2005), Christiano et al. (2007).

The two approaches can be combined (see, e.g., Gerlach and Smets (1995)).

Let us consider a simple example that could motivate short-run restrictions. Consider the following stylized macro model:

$$\begin{aligned} g_t &= \bar{g} - \lambda(i_{t-1} - \mathbb{E}_{t-1}\pi_t) + \underbrace{\sigma_d \eta_{d,t}}_{\text{demand shock}} && \text{(IS curve)} \\ \Delta\pi_t &= \beta(g_t - \bar{g}) + \underbrace{\sigma_\pi \eta_{\pi,t}}_{\text{cost push shock}} && \text{(Phillips curve)} \\ i_t &= \rho i_{t-1} + \left[\gamma_\pi \mathbb{E}_t \pi_{t+1} + \gamma_g (g_t - \bar{g})\right] \\ &\quad + \underbrace{\sigma_{mp} \eta_{mp,t}}_{\text{Mon. Pol. shock}} && \text{(Taylor rule),} \end{aligned} \tag{3.16}$$

where:

$$\eta_t = \begin{bmatrix} \eta_{\pi,t} \\ \eta_{d,t} \\ \eta_{mp,t} \end{bmatrix} \sim i.i.d.\, \mathcal{N}(0, I). \tag{3.17}$$

Vector η_t is assumed to be a vector of structural shocks, mutually and serially independent. On date t:

- g_t is contemporaneously affected by $\eta_{d,t}$ only;
- π_t is contemporaneously affected by $\eta_{\pi,t}$ and $\eta_{d,t}$;
- i_t is contemporaneously affected by $\eta_{mp,t}$, $\eta_{\pi,t}$ and $\eta_{d,t}$.

System (3.16) could be rewritten in the form:

$$\begin{bmatrix} d_t \\ \pi_t \\ i_t \end{bmatrix} = \Phi(L) \begin{bmatrix} d_{t-1} \\ \pi_{t-1} \\ i_{t-1}+ \end{bmatrix} + \underbrace{\underbrace{\begin{bmatrix} 0 & \bullet & 0 \\ \bullet & \bullet & 0 \\ \bullet & \bullet & \bullet \end{bmatrix}}_{=B} \eta_t}_{=\varepsilon_t} \tag{3.18}$$

This is the **reduced-form** of the model. This representation suggests three additional restrictions on the entries of B; the latter matrix is therefore identified (up to the signs of its columns) as soon as $\Omega = BB'$ is known.

There are particular cases in which some well-known matrix decomposition of $\Omega = \mathbb{V}ar(\varepsilon_t)$ can be used to easily estimate some specific SVAR.

Consider the following context:

- A first shock (say, $\eta_{n_1,t}$) can affect instantaneously (i.e., on date t) only one of the endogenous variable (say, $y_{n_1,t}$);
- A second shock (say, $\eta_{n_2,t}$) can affect instantaneously (i.e., on date t) two endogenous variables, $y_{n_1,t}$ (the same as before) and $y_{n_2,t}$;
- ...

This implies (1) that column n_1 of B has only 1 non-zero entry (this is the n_1^{th} entry), (2) that column n_2 of B has 2 non-zero entries (the n_1^{th} and the n_2^{th} ones), etc. Without loss of generality, we can set $n_1 = n$, $n_2 = n-1$, etc. In this context, matrix B is lower triangular.

The Cholesky decomposition of Ω_ε then provides an appropriate estimate of B, since this matrix decomposition yields to a lower triangular matrix satisfying:

$$\Omega_\varepsilon = BB'.$$

For instance, Dedola and Lippi (2005) estimate 5 structural VAR models for the US, the UK, Germany, France and Italy to analyse the monetary-policy transmission mechanisms. They estimate SVAR(5) models over the

period 1975-1997. The shock-identification scheme is based on Cholesky decompositions, the ordering of the endogenous variables being: the industrial production, the consumer price index, a commodity price index, the short-term rate, monetary aggregate and the effective exchange rate (except for the US). This ordering implies that monetary policy reacts to the shocks affecting the first three variables but that the latter react to monetary policy shocks with a one-period lag only.

Importantly, the Cholesky approach can be useful when one is interested in one specific structural shock. This was the case, e.g., of Christiano et al. (1996). Their identification is based on the following relationship between ε_t and η_t:

$$\begin{bmatrix} \varepsilon_{S,t} \\ \varepsilon_{r,t} \\ \varepsilon_{F,t} \end{bmatrix} = \begin{bmatrix} B_{SS} & 0 & 0 \\ B_{rS} & B_{rr} & 0 \\ B_{FS} & B_{Fr} & B_{FF} \end{bmatrix} \begin{bmatrix} \eta_{S,t} \\ \eta_{r,t} \\ \eta_{F,t} \end{bmatrix},$$

where S, r and F respectively correspond to *slow-moving variables*, the policy variable (short-term rate) and *fast-moving variables*. While $\eta_{r,t}$ is scalar, $\eta_{S,t}$ and $\eta_{F,t}$ may be vectors. The space spanned by $\varepsilon_{S,t}$ is the same as that spanned by $\eta_{S,t}$. As a result, because $\varepsilon_{r,t}$ is a linear combination of $\eta_{r,t}$ and $\eta_{S,t}$ (which are $\perp$), it comes that the $B_{rr}\eta_{r,t}$'s are the (population) residuals in the regression of $\varepsilon_{r,t}$ on $\varepsilon_{S,t}$. Because $\mathbb{V}ar(\eta_{r,t}) = 1$, B_{rr} is given by the square root of the variance of $B_{rr}\eta_{r,t}$. $B_{F,r}$ is finally obtained by regressing the components of $\varepsilon_{F,t}$ on the estimates of $\eta_{r,t}$.

An equivalent approach consists in computing the Cholesky decomposition of BB' and the contemporaneous impacts of the monetary policy shock (on the n endogenous variables) are the components of the column of B corresponding to the policy variable.

```
library(AEC)
library(vars)
data("USmonthly")
# Select sample period:
First.date <- "1965-01-01";Last.date <- "1995-06-01"
indic.first <- which(USmonthly$DATES==First.date)
indic.last  <- which(USmonthly$DATES==Last.date)
USmonthly   <- USmonthly[indic.first:indic.last,]
considered.variables <- c("LIP","UNEMP","LCPI","LPCOM","FFR","NBR","TTR","M1")
```

```
y <- as.matrix(USmonthly[considered.variables])
res.svar.ordering <- svar.ordering(y,p=3,
                                   posit.of.shock = 5,
                                   nb.periods.IRF = 20,
                                   nb.bootstrap.replications = 100,
                                   confidence.interval = 0.90, # expressed in pp.
                                   indic.plot = 1 # Plots are displayed if = 1.
)
```

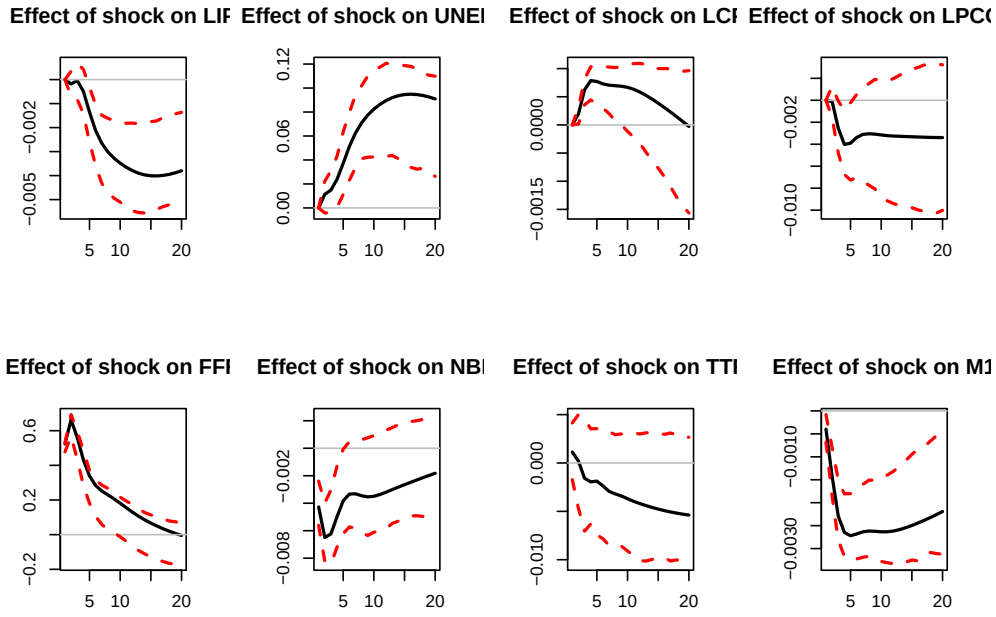


Figure 3.2: Response to a monetary-policy shock. Identification approach of Christiano, Eichenbaum and Evans (1996). Confidence intervals are obtained by boostrapping the estimated VAR model (see inference section).

Let us now turn to **Long-run restrictions**. Such a restriction concerns the long-run influence of a shock on an endogenous variable. Let us consider for instance a structural shock that is assumed to have no "long-run influence" on GDP. How to express this? The long-run change in GDP can be expressed as $GDP_{t+h} - GDP_t$, with h large. Note further that:

$$GDP_{t+h} - GDP_t = \Delta GDP_{t+h} + \Delta GDP_{t+h-1} + \cdots + \Delta GDP_{t+1}.$$

Hence, the fact that a given structural shock ($\eta_{i,t}$, say) has no long-run influence on GDP means that

$$\lim_{h\to\infty} \frac{\partial GDP_{t+h}}{\partial \eta_{i,t}} = \lim_{h\to\infty} \frac{\partial}{\partial \eta_{i,t}} \left(\sum_{k=1}^{h} \Delta GDP_{t+k} \right) = 0.$$

This can be easily formulated as a function of B and of the matrices Φ_i when y_t (including ΔGDP_t) follows a VAR process.

Without loss of generality, we will only consider the VAR(1) case. Indeed, one can always write a VAR(p) as a VAR(1). To see that, stack the last p values of vector y_t in vector $y_t^* = [y_t', ..., y_{t-p+1}']'$; Eq. (3.1) can then be rewritten in its **companion form**:

$$y_t^* = \underbrace{\begin{bmatrix} c \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{=c^*} + \underbrace{\begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_p \\ I & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & I & 0 \end{bmatrix}}_{=\Phi} y_{t-1}^* + \underbrace{\begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\varepsilon_t^*}, \tag{3.19}$$

where matrices Φ and $\Omega^* = \mathbb{V}ar(\varepsilon_t^*)$ are of dimension $np \times np$; Ω^* is filled with zeros, except the $n \times n$ upper-left block that is equal to $\Omega = \mathbb{V}ar(\varepsilon_t)$. (Matrix Φ had been introduced in Eq. (3.7).)

Focusing on the VAR(1) case:

$$\begin{aligned} y_t &= c + \Phi y_{t-1} + \varepsilon_t \\ &= c + \varepsilon_t + \Phi(c + \varepsilon_{t-1}) + ... + \Phi^k(c + \varepsilon_{t-k}) + ... \\ &= \mu + \varepsilon_t + \Phi\varepsilon_{t-1} + ... + \Phi^k\varepsilon_{t-k} + ... \\ &= \mu + B\eta_t + \Phi B\eta_{t-1} + ... + \Phi^k B\eta_{t-k} + ..., \end{aligned}$$

The sequence of shocks $\{\eta_t\}$ determines the sequence $\{y_t\}$. What if $\{\eta_t\}$ is replaced with $\{\tilde{\eta}_t\}$, where $\tilde{\eta}_t = \eta_t$ if $t \neq s$ and $\tilde{\eta}_s = \eta_s + \gamma$? Assume $\{\tilde{y}_t\}$ is the associated "perturbated" sequence. We have $\tilde{y}_t = y_t$ if $t < s$. For $t \geq s$, the Wold decomposition of $\{\tilde{y}_t\}$ implies:

$$\tilde{y}_t = y_t + \Phi^{t-s} B\gamma.$$

Therefore, the cumulative impact of γ on $\tilde{y}_t$ will be (for $t \geq s$):

$$\begin{aligned} (\tilde{y}_t - y_t) + (\tilde{y}_{t-1} - y_{t-1}) + \cdots + (\tilde{y}_s - y_s) \quad &= \\ (Id + \Phi + \Phi^2 + \cdots + \Phi^{t-s})B\gamma. \end{aligned} \tag{3.20}$$

Consider a shock on $\eta_{1,t}$, with a magnitude of 1. This shock corresponds to $\gamma = [1, 0, ..., 0]'$. Given Eq. (3.20), the long-run cumulative effect of this shock on the endogenous variables is given by:

$$\underbrace{(Id + \Phi + ... + \Phi^k + ...)}_{=(Id-\Phi)^{-1}} B \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

that is the first column of $\Theta \equiv (Id - \Phi)^{-1}B$.

In this context, consider the following long-run restriction: *"j^{th} structural shock has no cumulative impact on the i^{th} endogenous variable"*. It is equivalent to

$$\Theta_{ij} = 0,$$

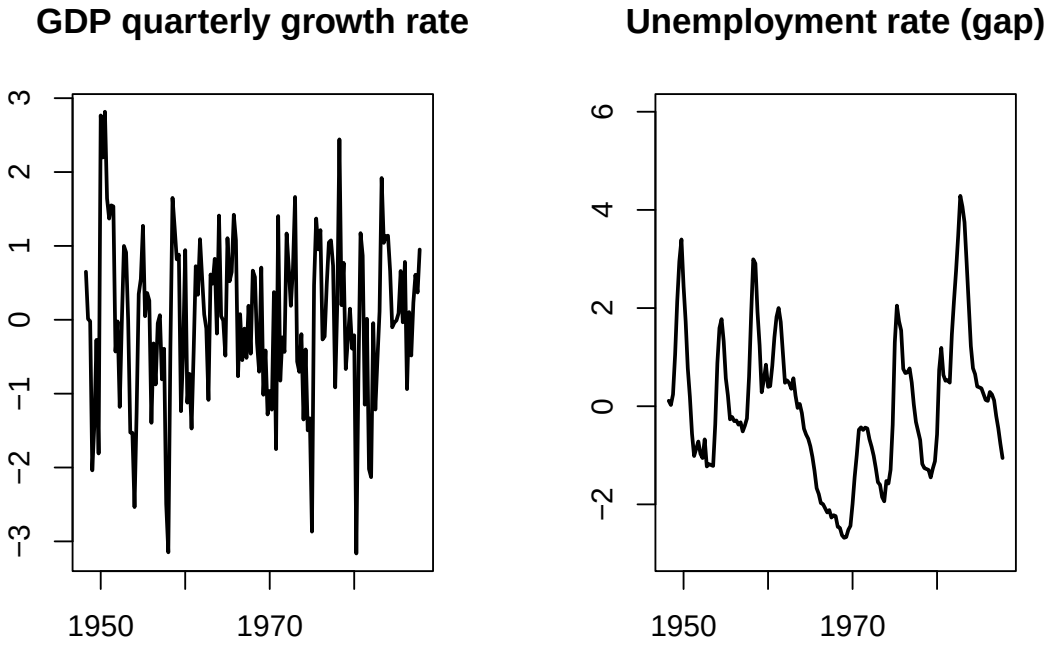
where Θ_{ij} is the element (i, j) of Θ.

Blanchard and Quah (1989) have implemented such long-run restrictions in a small-scale VAR. Two variables are considered: GDP and unemployment. Consequently, the VAR is affected by two types of shocks. Specifically, authors want to identify **supply shocks** (that can have a permanent effect on output) and **demand shocks** (that cannot have a permanent effect on output).[1]

Blanchard and Quah (1989)'s dataset is quarterly, spanning the period from 1950:2 to 1987:4. Their VAR features 8 lags. Here are the data they use:

```
library(AEC)
data(BQ)
par(mfrow=c(1,2))
plot(BQ$Date,BQ$Dgdp,type="l",main="GDP quarterly growth rate",
     xlab="",ylab="",lwd=2)
plot(BQ$Date,BQ$unemp,type="l",ylim=c(-3,6),main="Unemployment rate (gap)",
     xlab="",ylab="",lwd=2)
```

[1]The motivation of the authors regarding their long-run restrictions can be obtained from a traditional Keynesian view of fluctuations. The authors propose a variant of a model from Fischer (1977).



Estimate a reduced-form VAR(8) model:

```
library(vars)
y <- BQ[,2:3]
est.VAR <- VAR(y,p=8)
Omega <- var(residuals(est.VAR))
```

Now, let us define a loss function (`loss`) that is equal to zero if (a) $BB' = \Omega$ and (b) the element (1,1) of ΘB is equal to zero:

```
# Compute (Id - Phi)^{-1}:
Phi <- Acoef(est.VAR)
PHI <- make.PHI(Phi)
sum.PHI.k <- solve(diag(dim(PHI)[1]) - PHI)[1:2,1:2]
loss <- function(param){
  B <- matrix(param,2,2)
  X <- Omega - B %*% t(B)
  Theta <- sum.PHI.k[1:2,1:2] %*% B
  loss <- 10000 * ( X[1,1]^2 + X[2,1]^2 + X[2,2]^2 + Theta[1,1]^2 )
  return(loss)
```

```
}
res.opt <- optim(c(1,0,0,1),loss,method="BFGS",hessian=FALSE)
print(res.opt$par)
```

```
## [1]  0.8570358 -0.2396345  0.1541395  0.1921221
```

(Note: one can use that type of approach, based on a loss function, to mix short- and long-run restrictions.)

Figure 3.3 displays the resulting IRFs. Note that, for GDP, we cumulate the GDP growth IRF, so as to have the response of the GDP in level.

```
B.hat <- matrix(res.opt$par,2,2)
print(cbind(Omega,B.hat %*% t(B.hat)))
```

```
##              Dgdp       unemp
## Dgdp   0.7582704 -0.17576173  0.7582694 -0.17576173
## unemp -0.1757617  0.09433658 -0.1757617  0.09433558
```

```
nb.sim <- 40
par(mfrow=c(2,2));par(plt=c(.15,.95,.15,.8))
Y <- simul.VAR(c=matrix(0,2,1),Phi,B.hat,nb.sim,y0.star=rep(0,2*8),
               indic.IRF = 1,u.shock = c(1,0))
plot(cumsum(Y[,1]),type="l",lwd=2,xlab="",ylab="",main="Demand shock on GDP")
plot(Y[,2],type="l",lwd=2,xlab="",ylab="",main="Demand shock on UNEMP")
Y <- simul.VAR(c=matrix(0,2,1),Phi,B.hat,nb.sim,y0.star=rep(0,2*8),
               indic.IRF = 1,u.shock = c(0,1))
plot(cumsum(Y[,1]),type="l",lwd=2,xlab="",ylab="",main="Supply shock on GDP")
plot(Y[,2],type="l",lwd=2,xlab="",ylab="",main="Supply shock on UNEMP")
```

3.0.7 Inference

Consider the following SVAR model:

$$y_t = \Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p} + \varepsilon_t$$

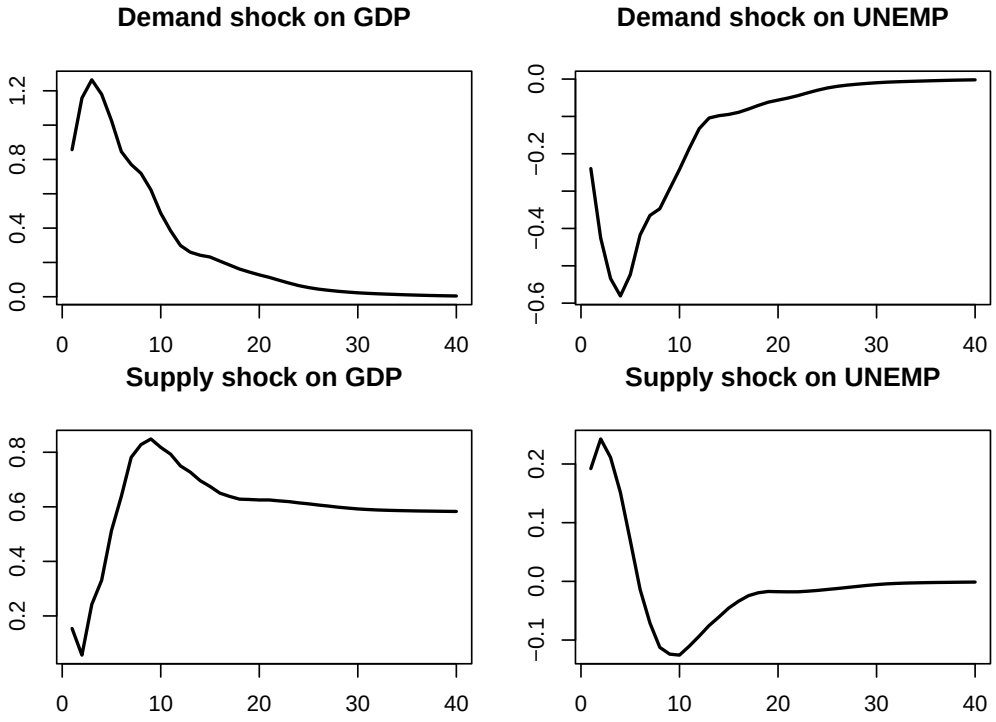


Figure 3.3: IRF of GDP and unemployment to demand and supply shocks.

with $\varepsilon_t = B\eta_t$, $\Omega_\varepsilon = BB'$.

The corresponding infinite MA representation (Eq. (3.3), or Wold theorem, Theorem 2.2) is:

$$y_t = \sum_{h=0}^{\infty} \Psi_h \eta_{t-h},$$

where $\Psi_0 = B$ and for $h = 1, 2, ...$:

$$\Psi_h = \sum_{j=1}^{h} \Psi_{h-j} \Phi_j,$$

with $\Phi_j = 0$ for $j > p$ (see Prop. 2.7 for this recursive computation of the Ψ_j's).

Inference on the VAR coefficients $\{\Phi_j\}_{j=1,...,p}$ is straightforward (standard OLS inference). But inference is more complicated regarding IRF. Indeed, as shown by the previous equation, the (infinite) MA coefficients $\{\Psi_j\}_{j=1,...}$ are non-linear functions of the $\{\Phi_j\}_{j=1,...,p}$ and of Ω_ε. An other issue pertain to small sample bias: typically, for persistent process, auto-regressive parameters are known to be downward biased.

The main inference methods are the following:

- Monte Carlo method (Hamilton (1994))
- Asymptotic normal approximation (Lütkepohl (1990)), or Delta method
- Bootstrap method (Kilian_1998)

Monte Carlo method

We use Monte Carlo when we need to approximate the distribution of a variable whose distribution is unknown (here: the Ψ_j's) but which is a function of another variable whose distribution is known (here, the Φ_j's).

For instance, suppose we know the distribution of a random variable X, which takes values in $\mathbb{R}$, with density function p. Assume we want to compute the mean of $\varphi(X)$. We have:

$$\mathbb{E}(\varphi(X)) = \int_{-\infty}^{+\infty} \varphi(x)p(x)dx$$

Suppose that the above integral does not have a simple expression. We cannot compute $\mathbb{E}(\varphi(X))$ but, by virtue of the law of large numbers (Theorem **??**), we can approximate it as follows:

$$\mathbb{E}(\varphi(X)) \approx \frac{1}{N}\sum_{i=1}^{N}\varphi(X^{(i)}),$$

where $\{X^{(i)}\}_{i=1,\dots,N}$ are N independent draws of X. More generally, the distribution of $\varphi(X)$ can be approximated by the empirical distribution of the $\varphi(X^{(i)}$'s. Typically, if 10'000 values of $\varphi(X^{(i)}$ are drawn, the 5^{th} percentile of the p.d.f. of $\varphi(X)$ can be approximated by the 500^{th} value of the 10'000 draws of $\varphi(X^{(i)}$ (after arranging these values in ascending order).

As regards the computation of confidence intervals around IRFs, one has to think of $\{\widehat{\Phi}_j\}_{j=1,\dots,p}$, and of $\widehat{\Omega}$ as X and $\{\widehat{\Psi}_j\}_{j=1,\dots}$ as $\varphi(X)$. (Proposition 3.3 provides us with the asymptotic distribution of the "X.")

To summarize, here are the steps one can implement to derive confidence intervals for the IRFs using the Monte-Carlo approach:

For each iteration k:

1. Draw $\{\widehat{\Phi}_j^{(k)}\}_{j=1,\dots,p}$ and $\widehat{\Omega}^{(k)}$ from their asymptotic distribution (using Proposition 3.3).
2. Compute the matrix $B^{(k)}$ so that $\widehat{\Omega}^{(k)} = B^{(k)}B^{(k)'}$, according to your identification strategy.
3. Compute the associated IRFs $\{\widehat{\Psi}_j\}^{(k)}$.

Perform N replications and report the median impulse response (and its confidence intervals).

Delta method

Suppose β is a vector of parameters and β is an estimator such that

$$\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} \mathcal{N}(0,\Sigma_\beta),$$

where d denotes convergence in distribution, $N(0,\Sigma_\beta)$ denotes the multivariate normal distribution with mean vector 0 and covariance matrix Σ_β and T is the sample size used for estimation.

Let $g(\beta) = (g_l(\beta), ..., g_m(\beta))'$ be a continuously differentiable function with values in $\mathbb{R}^m$, and assume that $\partial g_i/\partial\beta' = (\partial g_i/\partial\beta_j)$ is nonzero at β for $i = 1, ..., m$. Then

$$\sqrt{T}(g(\hat{\beta}) - g(\beta)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial g}{\partial \beta'}\Sigma_\beta \frac{\partial g'}{\partial \beta}\right).$$

(This formula underlies the Delta method, see Eq. (**??**).)

Using this property, Lütkepohl (1990) provides the asymptotic distributions of the Ψ_j's.

A limit of the last two approaches (Monte Carlo and the Delta method) is that they critically rely on asymptotic results. Boostrapping approaches are more robust in small-sample situations.

Bootstrap

IRFs' confidence intervals are intervals where 90% (or 95%,75%,...) of the IRFs would lie, if we were to repeat the estimation a large number of times in similar conditions (T observations). We obviously cannot do this, because we have only one sample: $\{y_t\}_{t=1,..,T}$. But we can try to *construct* such samples.

Bootstrapping consists in:

- re-sampling N times, i.e., constructing N samples of T observations, using the estimated VAR coefficients and

a. a sample of residuals from the distribution $N(0, BB')$ (**parametric approach**), or
b. a sample of residuals drawn randomly from the set of the actual estimated residuals $\{\hat{\varepsilon}_t\}_{t=1,..,T}$. (**non-parametric approach**).

- re-estimating the SVAR N times.

Here is the algorithm:

1. Construct a sample

$$y_t^{(k)} = \widehat{\Phi}_1 y_{t-1}^{(k)} + \cdots + \widehat{\Phi}_p y_{t-p}^{(k)} + \hat{\varepsilon}_t^{(k)},$$

with $\hat{\varepsilon}_t^{(k)} = \hat{\varepsilon}_{s_t^{(k)}}$, where $\{s_1^{(k)}, .., s_T^{(k)}\}$ is a random set from $\{1, .., T\}^T$.

2. Re-estimate the SVAR and compute the IRFs $\{\widehat{\Psi}_j\}^{(k)}$.

Perform N replications and report the median impulse response (and its confidence intervals).

Bootstrap-after-bootstrap (Kilian (1998))

The previous simple bootstrapping procedure deals with non-normality and small sample distribution, since we use the actual residuals. However, it does not deal with the *small sample bias*, stemming, in particular, from small-sample bias associated with OLS coefficient estimates $\{\widehat{\Phi}_j\}_{j=1,..,p}$. The main idea of the bootstrap-after-bootstrap of Kilian (1998) is to run two consecutive boostraps: the objective of the first is to compute the bias, which can further be used to correct the initial estimates of the Φ_i's. Further, these corrected estimates are used —in the second boostrap— to compute a set of IRFs (as in the standard boostrap).

More formally, the algorithm is as follows:

1. Estimate the SVAR coefficients $\{\widehat{\Phi}_j\}_{j=1,..,p}$ and $\widehat{\Omega}$
2. **First bootstrap.** For each iteration k:

a. Construct a sample

$$y_t^{(k)} = \widehat{\Phi}_1 y_{t-1}^{(k)} + \cdots + \widehat{\Phi}_p y_{t-p}^{(k)} + \hat{\varepsilon}_t^{(k)},$$

with $\hat{\varepsilon}_t^{(k)} = \hat{\varepsilon}_{s_t^{(k)}}$, where $\{s_1^{(k)}, .., s_T^{(k)}\}$ is a random set from $\{1, .., T\}^T$.

b. Re-estimate the VAR and compute the coefficients $\{\widehat{\Phi}_j\}_{j=1,..,p}^{(k)}$.

3. Perform N replications and compute the median coefficients $\{\widehat{\Phi}_j\}_{j=1,..,p}^{*}$.
4. Approximate the bias terms by $\widehat{\Theta}_j = \widehat{\Phi}_j^{*} - \widehat{\Phi}_j$.
5. Construct the bias-corrected terms $\widetilde{\Phi}_j = \widehat{\Phi}_j - \widehat{\Theta}_j$.
6. **Second bootstrap.** For each iteration k:

a. Construct a sample now from

$$y_t^{(k)} = \widetilde{\Phi}_1 y_{t-1}^{(k)} + \cdots + \widetilde{\Phi}_p y_{t-p}^{(k)} + \hat{\varepsilon}_t^{(k)}.$$

b. Re-estimate the VAR and compute the coefficients $\{\widehat{\Phi}_j^*\}_{j=1,..,p}^{(k)}$.
c. Construct the bias-corrected estimates $\widetilde{\Phi}_j^{*(k)} = \widehat{\Phi}_j^{*(k)} - \widehat{\Theta}_j$.
d. Compute the associated IRFs $\{\widetilde{\Psi}_j^{*(k)}\}_{j\geq 1}$.

7. Perform N replications and compute the median and the confidence interval of the set of IRFs.

It should be noted that correcting for the bias can generate non-stationary results ($\tilde{\Phi}$ with eigenvalue with modulus > 1). Solution (Kilian (1998)):

In step 5, check if the largest eigenvalue of $\tilde{\Phi}$ is of modulus <1. If not, shrink the bias: for all js, set $\widehat{\Theta}_j^{(i+1)} = \delta_{i+1}\widehat{\Theta}_j^{(i)}$, with $\delta_{i+1} = \delta_i - 0.01$, starting with $\delta_1 = 1$ and $\widehat{\Theta}_j^{(1)} = \widehat{\Theta}_j$, and compute $\widetilde{\Phi}_j^{(i+1)} = \widehat{\Phi}_j - \widehat{\Theta}_j^{(i+1)}$ until the largest eigenvalue of $\tilde{\Phi}^{(i+1)}$ has modulus <1.

Function `VAR.Boot` of package `VAR.etp` (Kim (2022)) can be used to operate the bias-correction approach of Kilian (1998):

```
library(VAR.etp)
library(vars) #standard VAR models
data(dat) # part of VAR.etp package
corrected <- VAR.Boot(dat,p=2,nb=200,type="const")
noncorrec <- VAR(dat,p=2)
rbind(corrected$coef[1,],
      (corrected$coef+corrected$Bias)[1,],
      noncorrec$varresult$inv$coefficients)
```

```
##         inv(-1)   inc(-1)   con(-1)    inv(-2)    inc(-2)   con(-2)       const
## [1,] -0.3160728 0.1718247 0.9606726 -0.1351475 0.08833738 0.9665836 -0.01791907
## [2,] -0.3196310 0.1459888 0.9612190 -0.1605511 0.11460498 0.9343938 -0.01672199
## [3,] -0.3196310 0.1459888 0.9612190 -0.1605511 0.11460498 0.9343938 -0.01672199
```

Chapter 4

Forecasting

Forecasting has always been an important part of the time series field (De Gooijer and Hyndman (2006)). Macroeconomic forecasts are done in many places: Public Administration (notably Treasuries), Central Banks, International Institutions (e.g. IMF, OECD), banks, big firms. These institutions are interested in the **point estimates** ($\sim$ most likely value) of the variable of interest. They also sometimes need to measure the **uncertainty** ($\sim$ dispersion of likely outcomes) associated to the point estimates.[1]

Forecasts produced by professional forecasters are available on these web pages:

- Philly Fed Survey of Professional Forecasters.
- ECB Survey of Professional Forecasters.
- IMF World Economic Outlook.
- OECD Global Economic Outlook.
- European Commission Economic Forecasts.

How to formalize the forecasting problem? Assume the current date is t. We want to forecast the value that variable y_t will take on date $t+1$ (i.e., y_{t+1}) based on the observation of a set of variables gathered in vector x_t (x_t may contain lagged values of y_t).

[1] In its inflation report, the Bank of England displays charts showing the conditional distribution of future inflation, called fan charts. This fan charts show the uncertainty associated with future inflation. See this page.

The forecaster aims at minimizing (a function of) the forecast error. It is usal to consider the following (quadratic) loss function:

$$\underbrace{\mathbb{E}([y_{t+1} - y_{t+1}^*]^2)}_{\text{Mean square error (MSE)}}$$

where y_{t+1}^* is the forecast of y_{t+1} (function of x_t).

Proposition 4.1 (Smallest MSE)**.** *The smallest MSE is obtained with MSEthe expectation of y_{t+1} conditional on x_t.*

Proof. See Appendix 8.5. □

Proposition 4.2. *Among the class of linear forecasts, the smallest MSE is obtained with the linear projection of y_{t+1} on x_t. This projection, denoted by $\widehat{P}(y_{t+1}|x_t) := \alpha' x_t$, satisfies:*

$$\mathbb{E}\left([y_{t+1} - \alpha' x_t]x_t\right) = \mathbf{0}. \tag{4.1}$$

Proof. Consider the function $f : \alpha \to \mathbb{E}\left([y_{t+1} - \alpha' x_t]^2\right)$. We have:

$$f(\alpha) = \mathbb{E}\left(y_{t+1}^2 - 2y_t x_t'\alpha + \alpha' x_t x_t'\alpha]\right).$$

We have $\partial f(\alpha)/\partial\alpha = \mathbb{E}(-2y_{t+1}x_t + 2x_t x_t'\alpha)$. The function is minimised for $\partial f(\alpha)/\partial\alpha = 0$. □

Eq. (4.1) implies that $\mathbb{E}\left(y_{t+1}x_t\right) = \mathbb{E}\left(x_t x_t'\right)\alpha$. (Note that $x_t x_t'\alpha = x_t(x_t'\alpha) = (\alpha' x_t)x_t'$.)

Hence, if $\mathbb{E}\left(x_t x_t'\right)$ is nonsingular,

$$\alpha = \left[\mathbb{E}\left(x_t x_t'\right)\right]^{-1}\mathbb{E}\left(y_{t+1}x_t\right). \tag{4.2}$$

The MSE then is:

$$\mathbb{E}([y_{t+1} - \alpha' x_t]^2) = \mathbb{E}(y_{t+1}^2) - \mathbb{E}\left(y_{t+1}x_t'\right)\left[\mathbb{E}\left(x_t x_t'\right)\right]^{-1}\mathbb{E}\left(x_t y_{t+1}\right).$$

Consider the regression $y_{t+1} = \beta' \mathbf{x}_t + \varepsilon_{t+1}$. The OLS estimate is:

$$\mathbf{b} = \left[\underbrace{\frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_t\mathbf{x}_t'}_{\mathbf{m}_1}\right]^{-1}\left[\underbrace{\frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_t' y_{t+1}}_{\mathbf{m}_2}\right].$$

If $\{x_t, y_t\}$ is covariance-stationary and ergodic for the second moments then the sample moments ($\mathbf{m}_1$ and $\mathbf{m}_2$) converges in probability to the associated population moments and $\mathbf{b} \xrightarrow{p} \alpha$ (where α is defined in Eq. (4.2)).

Example 4.1 (Forecasting an MA(q) process)**.** Consider the MA(q) process:

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\{\varepsilon_t\}$ is a white noise sequence (Def. 1.1).

We have:

$$\mathbb{E}(y_{t+h}|\varepsilon_t, \varepsilon_{t-1}, \dots) = \\ \begin{cases} \mu + \theta_h \varepsilon_t + \cdots + \theta_q \varepsilon_{t-q+h} & for \quad h \in [1, q] \\ \mu & for \quad h > q \end{cases}$$

and

$$\mathbb{V}ar(y_{t+h}|\varepsilon_t, \varepsilon_{t-1}, \dots) = \mathbb{E}\left([y_{t+h} - \mathbb{E}(y_{t+h}|\varepsilon_t, \varepsilon_{t-1}, \dots)]^2\right) = \\ \begin{cases} \sigma^2(1 + \theta_1^2 + \cdots + \theta_{h-1}^2) & for \quad h \in [1, q] \\ \sigma^2(1 + \theta_1^2 + \cdots + \theta_q^2) & for \quad h > q. \end{cases}$$

Remark: The previous reasoning relies on the assumption that the ε_ts are observed. But this is generally not the case in practice. Note that consistent estimates are available if the MA process is invertible (see Eq. (2.26)) .

Example 4.2 (Forecasting an AR(p) process)**.** (See this web interface.) Consider the AR(p) process:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ is a white noise sequence (Def. 1.1).

Using the notation of Eq. (2.3), we have:

$$\mathbf{y}_t - \mu = F(\mathbf{y}_{t-1} - \mu) + \xi_t,$$

with $\mu = [\mu, \dots, \mu]'$ (μ is defined in Eq. (2.8)). Hence:

$$\mathbf{y}_{t+h} - \mu = \xi_{t+h} + F\xi_{t+h-1} + \cdots + F^{h-1}\xi_{t+1} + F^h(\mathbf{y}_t - \mu).$$

Therefore:

$$\begin{aligned} \mathbb{E}(\mathbf{y}_{t+h}|y_t, y_{t-1}, \dots) &= \mu + F^h(\mathbf{y}_t - \mu) \\ \mathbb{V}ar\left([\mathbf{y}_{t+h} - \mathbb{E}(\mathbf{y}_{t+h}|y_t, y_{t-1}, \dots)]\right) &= \Sigma + F\Sigma F' + \cdots + F^{h-1}\Sigma(F^{h-1})', \end{aligned}$$

where:

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & \dots \\ 0 & 0 & \\ \vdots & & \ddots \end{bmatrix}.$$

Alternative approach: Taking the (conditional) expectations of both sides of

$$y_{t+h} - \mu = \phi_1(y_{t+h-1} - \mu) + \phi_2(y_{t+h-2} - \mu) + \cdots + \phi_p(y_{t-p} - \mu) + \varepsilon_{t+h},$$

we obtain:

$$\begin{aligned} \mathbb{E}(y_{t+h}|y_t, y_{t-1}, \dots) &= \mu + \phi_1\left(\mathbb{E}[y_{t+h-1}|y_t, y_{t-1}, \dots] - \mu\right) + \\ &\quad \phi_2\left(\mathbb{E}[y_{t+h-2}|y_t, y_{t-1}, \dots] - \mu\right) + \cdots + \\ &\quad \phi_p\left(\mathbb{E}[y_{t+h-p}|y_t, y_{t-1}, \dots] - \mu\right), \end{aligned}$$

which can be exploited recursively.

The recursion begins with $\mathbb{E}(y_{t-k}|y_t, y_{t-1}, \dots) = y_{t-k}$ (for any $k \geq 0$).

Example 4.3 (Forecasting an ARMA(p,q) process)**.** Consider the process:

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}, \tag{4.3}$$

where $\{\varepsilon_t\}$ is a white noise sequence (Def. 1.1). We assume that the MA part of the process is invertible (see Eq. (2.26)), which implies that the information contained in $\{y_t, y_{t-1}, y_{t-2}, \dots\}$ is identical to that in $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots\}$.

While one could use a recursive algorithm to compute the conditional mean (as in Example 4.2), it is convenient to employ the Wold decomposition of this process (see Theorem 2.2 and Prop. 2.7 for the computation of the ψ_i's in the context of ARMA processes):

$$y_t = \mu + \sum_{i=0}^{+\infty} \psi_i \varepsilon_{t-i}.$$

This implies:

$$\begin{aligned} y_{t+h} &= \mu + \sum_{i=0}^{h-1} \psi_i \varepsilon_{t+h-i} + \sum_{i=h}^{+\infty} \psi_i \varepsilon_{t+h-i} \\ &= \mu + \sum_{i=0}^{h-1} \psi_i \varepsilon_{t+h-i} + \sum_{i=0}^{+\infty} \psi_{i+h} \varepsilon_{t-i}. \end{aligned}$$

Since $\mathbb{E}(y_{t+h}|y_t, y_{t-1}, \dots) = \mu + \sum_{i=0}^{+\infty} \psi_{i+h}\varepsilon_{t-i}$, we get:

$$\mathbb{V}ar(y_{t+h}|y_t, y_{t-1}, \dots) = \mathbb{V}ar\left(\sum_{i=0}^{h-1} \psi_i \varepsilon_{t+h-i}\right) = \sigma^2 \sum_{i=0}^{h-1} \psi_i^2.$$

How to use the previous formulas in practice?

One has first to select a specification and to estimate the model. Two methods to determine relevant specifications:

a. Information criteria (see Definition 2.11).
b. Box-Jenkins approach.

Box and Jenkins (1976) have proposed an approach that is now widely used.

1. Data transformation. The data should be transformed to "make them stationary". To do so, one can e.g. take logarithms, take changes in the considered series, remove (deterministic) trends.
2. Select p and q. This can be based on the PACF approach (see Section 2.0.4), or on selection criteria (see Definition 2.11).
3. Estimate the model parameters. See Section 2.0.8.
4. Check that the estimated model is consistent with the data. See below.

Assessing the performances of a forecasting model

Once one has fitted a model on a given dataset (of length T, say), one compute MSE (mean square errors) to evaluate the performance of the model. But this MSE is the **in-sample** one. It is easy to reduce in-sample MSE. Typically, if the model is estimated by OLS, adding covariates mechanically

reduces the MSE (see Props. **??** and **??**). That is, even if additional data are irrelevant, the R^2 of the regression increases. Adding irrelevant variables increases the (in-sample) R^2 but is bound to increase the **out-of-sample** MSE.

Therefore, it is important to analyse **out-of-sample** performances of the forecasting model:

a. Estimate a model on a sample of reduced size $(1, ..., T^*$, with $T^* < T)$
b. Use the remaining available periods $(T^* + 1, ..., T)$ to compute **out-of-sample** forecasting errors (and compute their MSE). In an out-of-sample exercise, it is important to make sure that the data used to produce a forecasts (as of date T^*) where indeed available on date T^*.

Diebold-Mariano test

How to compare different forecasting approaches? Diebold and Mariano (1995) have proposed a simple test to address this question.

Assume that you want to compare approaches A and B. You have historical data sets and you have implemented both approaches in the past, providing you with two sets of forecasting errors: $\{e_t^A\}_{t=1,...,T}$ and $\{e_t^B\}_{t=1,...,T}$.

It may be the case that your forecasts serve a specific purpose and that, for instance, you dislike positive forecasting errors and you care less about negative errors. We assume you are able to formalise this by means of a **loss function** $L(e)$. For instance:

- If you dislike large positive errors, you may set $L(e) = \exp(e)$.
- If you are concerned about both positive and negative errors (indifferently), you may set $L(e) = e^2$ (standard approach).

Let us define the sequence $\{d_t\}_{t=1,...,T} \equiv \{L(e_t^A) - L(e_t^B)\}_{t=1,...,T}$ and assume that this sequence is covariance stationary. We consider the following null hypothesis: $H_0 : \bar{d} = 0$, where $\bar{d}$ denotes the population mean of the d_ts. Under H_0 and under the assumption of covariance-stationarity of d_t, we have (Theorem @ref{(hm:CLTcovstat)):

$$\sqrt{T}\bar{d}_T \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=-\infty}^{+\infty} \gamma_j\right),$$

where the γ_js are the autocovariances of d_t.

Hence, assuming that $\hat{\sigma}^2$ is a consistent estimate of $\sum_{j=-\infty}^{+\infty} \gamma_j$ (for instance the one given by the Newey-West formula, see Def. **??**), we have, under H_0:

$$DM_T := \sqrt{T}\frac{\bar{d}_T}{\sqrt{\hat{\sigma}^2}} \xrightarrow{d} \mathcal{N}(0,1).$$

DM_T is the test statistics. For a test of size α, the critical region is:[2]

$$]-\infty, -\Phi^{-1}(1-\alpha/2)] \cup [\Phi^{-1}(1-\alpha/2), +\infty[,$$

where Φ is the c.d.f. of the standard normal distribution.

Example 4.4 (Forecasting Swiss GDP growth)**.** We use a long historical time series of the Swiss GDP growth taken from the Jordà et al. (2017) dataset (see Figure 1.3, and Example 2.4).

We want to forecast this GDP growth. We envision two specifications : an AR(1) specification (the one advocated by the AIC criteria, see Example 2.4), and an ARMA(2,2) specification. We are interested in 2-year-ahead forecasts (i.e., $h = 2$ since the data are yearly).

```
library(AEC)
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
data <- subset(JST,iso=="CHE")
T <- dim(data)[1]
y <- c(NaN,log(data$gdp[2:T]/data$gdp[1:(T-1)]))
first.date <- T-50
e1 <- NULL; e2 <- NULL;h<-2
```

[2]This ShinyApp application illustrates the notion of statistical test (illustrating the p-value and the cirtical region, in particular).

```
for(T.star in first.date:(T-h)){
  estim.model.1 <- arima(y[1:T.star],order=c(1,0,0))
  estim.model.2 <- arima(y[1:T.star],order=c(2,0,2))
  e1 <- c(e1,y[T.star+h] - predict(estim.model.1,n.ahead=h)$pred[h])
  e2 <- c(e2,y[T.star+h] - predict(estim.model.2,n.ahead=h)$pred[h])
}
res.DM <- dm.test(e1,e2,h = h,alternative = "greater")
res.DM
```

```
##
##  Diebold-Mariano Test
##
## data:  e1e2
## DM = -0.82989, Forecast horizon = 2, Loss function power = 2, p-value =
## 0.7946
## alternative hypothesis: greater
```

With `alternative = "greater"` The alternative hypothesis is that method 2 is more accurate than method 1. Since we do not reject the null (the p-value being of 0.795), we are not led to use the more sophisticated model (ARMA(2,2)) and we keep the simple AR(1) model.

Assume now that we want to compare the AR(1) process to a VAR model (see Def. 3.1). We consider a bivariate VAR, where GDP growth is complemented with CPI-based inflation rate.

```
library(vars)
infl <- c(NaN,log(data$cpi[2:T]/data$cpi[1:(T-1)]))
y_var <- cbind(y,infl)
e3 <- NULL
for(T.star in first.date:(T-h)){
  estim.model.3 <- VAR(y_var[2:T.star,],p=1)
  e3 <- c(e3,y[T.star+h] - predict(estim.model.3,n.ahead=h)$fcst$y[h,1])
}
res.DM <- dm.test(e1,e2,h = h,alternative = "greater")
res.DM
```

```
##
##  Diebold-Mariano Test
##
## data:  e1e2
## DM = -0.82989, Forecast horizon = 2, Loss function power = 2, p-value =
## 0.7946
## alternative hypothesis: greater
```

Again, we do not find that the alternative model (here the VAR(1) model) is better than the AR(1) model to forecast GDP growth.

Chapter 5

Non-stationary processes

In time series analysis, nonstationarity has several crucial implications. Indeed, various time-series regression procedures are not reliable anymore when processes are nonstationary.

There are different reasons why a process can be nonstationary. Two examples are trends and breaks. Generally speaking, a trend is a persistent long-term movement of a variable over time. A linear trend is a simple example of (deterministic) trend. We say that process y_t is stationary around a linear trend if it is given by:

$$y_t = a + bt + x_t,$$

where x_t is a stationary process. But "trends" may also be stochastic. A typical example of stochastic trend is a random walk:

$$w_t = w_{t-1} + \varepsilon_t,$$

where ε_t is a sequence of i.i.d. mean-zero shocks with variance σ^2.

As we will see, if $y_t = w_t + x_t$, where w_t is a random walk, then the use of y_t in econometric specifications will have to be operated carefully. Typically, as we shall see, standard inference in linear regression models may no longer be valid since y_t features no unconditional moments.

We have

$$\mathbb{V}ar_t(y_{t+h}) = \mathbb{V}ar_t(\varepsilon_{t+h}) + \cdots + \mathbb{V}ar_t(\varepsilon_{t+1}) = h\sigma^2.$$

Using the law of total variance (and assuming that $\mathbb{V}ar(y_t)$ exists), we have, for any h:

$$\mathbb{V}ar(y_t) = \mathbb{E}[\mathbb{V}ar_{t-h}(y_t)] + \mathbb{V}ar[\mathbb{E}_{t-h}(y_t)]. \tag{5.1}$$

We also have $\mathbb{E}_{t-h}(y_t) = y_{t-h}$ and $\mathbb{V}ar_t(y_{t+h}) = \mathbb{V}ar_t(\varepsilon_{t+h}) + \cdots + \mathbb{V}ar_t(\varepsilon_{t+1}) = h\sigma^2$. Therefore, Eq. (5.1) gives:

$$\mathbb{V}ar(y_t) = h\sigma^2 + \mathbb{V}ar(y_{t-h}).$$

Since the previous equation needs to be satisfied for any h (even infintely large ones), and since $\mathbb{V}ar(y_{t-h}) > 0$, $\mathbb{V}ar(y_t)$ can not be finite. None of the population moments of a random walk is actually defined.

5.1 Issues when working with nonstationary time series

Let us discuss three issues associated with nonstationary processes: 1. Bias of autoregressive coefficient towards 0 (context: autoregressions). 2. Even in large samples, the t-statistics cannot be approximated by the normal distribution (context: OLS regressions). 3. Spurious regressions: OLS-based regression analysis tends to indicate relationships between *independent* nonstationary (or unit-root) series.

5.1.1 The bias of autoregressive coefficient towards zero

Consider a non-stationary y_t process (e.g., y_t follows a random walk). The estimate of ϕ in the (OLS) regression $y_t = c + \phi y_{t-1} + \varepsilon_t$ is then biased toward zero. In particular, we have the approximation $\mathbb{E}(\phi) = 1 - 5.3/T$, where T is the sample size (see, e.g., Abadir (1993)). The 5th percentile of the distribution of ϕ is approximately $1-14.1/T$, e.g., 0.824 for $T = 80$. This notably poses problems for forecasts. Indeed, while we have $\mathbb{E}_t(y_{t+h}) = y_t$ if y_t follows a random walk, someone who would fit an AR(1) process and would obtain for instance $\widehat{\phi} = 0.824$ would obtain that $\mathbb{E}_t(y_{t+10}) = 0.144y_t$.

This is illustrated by Figure 5.1. This figure shows, in black, the distributions of $\widehat{\phi}$, obtained by OLS in the context of the linear model $y_t = c + \phi y_{t-1} + \varepsilon_t$.

These distributions are obtained by simulations. We consider two sample sizes ($T = 50$, left plot, and $T = 200$, right plot). The vertical red bars indicate the means of the distributions, the vertical blue bar gives the approximated mean given by $1 - 5.3/T$.

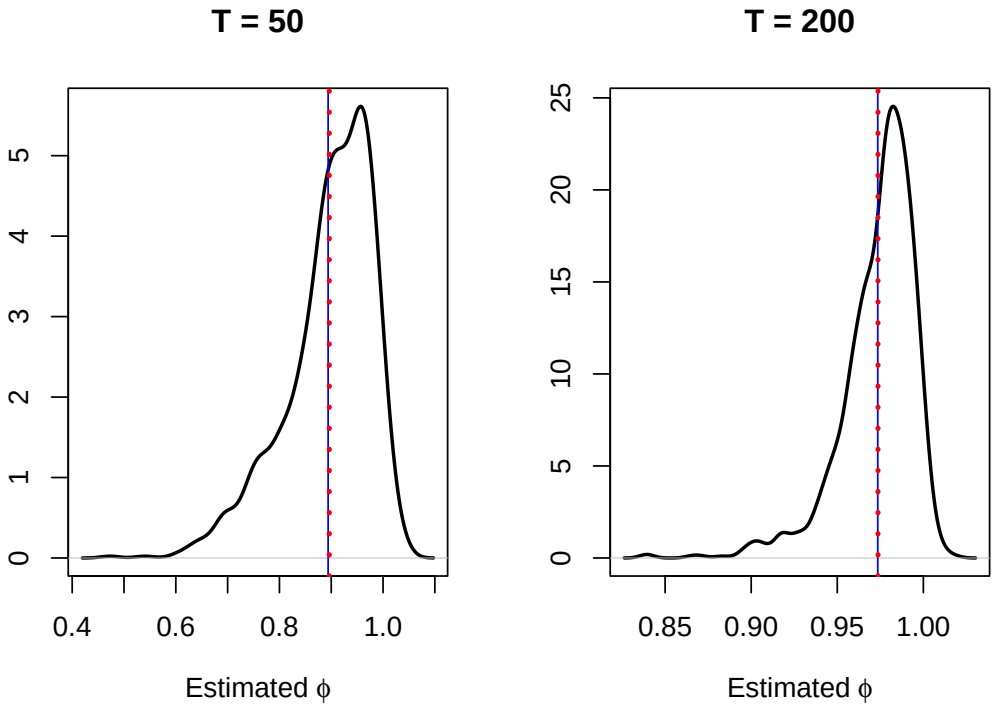


Figure 5.1: The densities of based on 1000 simulations of samples of length T, they are approximated by the kernel approach.

5.1.2 Spurious regressions

Consider two independent non-stationary (unit roots) variables. If we regress one on the other, and if we use the standard OLS formulas to compute the standard deviation of the regression parameter, it can be shown that we will tend to underestimate this standard deviation. As a result, if we use standard t-statistics to test for the existence of a relationship between the two variables, we will often have false positive, i.e., we will often reject the null hypothesis of no relationship (H_0: regression coefficient = 0). This phenomenon is called *spurious regressions* (see, e.g., these examples).

This situation is illustrated by Figure 5.2. We simulate two independent random walks, x_t and y_t, and we regress y_t on x_t by OLS. (We consider two different sample sizes, $T = 50$ and $T = 200$.) The black lines represent the densities of the estimated β's, that are the slope coefficients of the regressions. The blue density represent the asymptotic distribution of β based on to the standard OLS formula (normal distribution of mean zero and of variance $\hat{\sigma}^2/\widehat{\mathbb{V}ar}(x_t)$). The fact that the latter distribution features a smaller variance than the former implies that using the standard OLS inference is misleading, as this distribution overestimates the accuracy of the estimator.

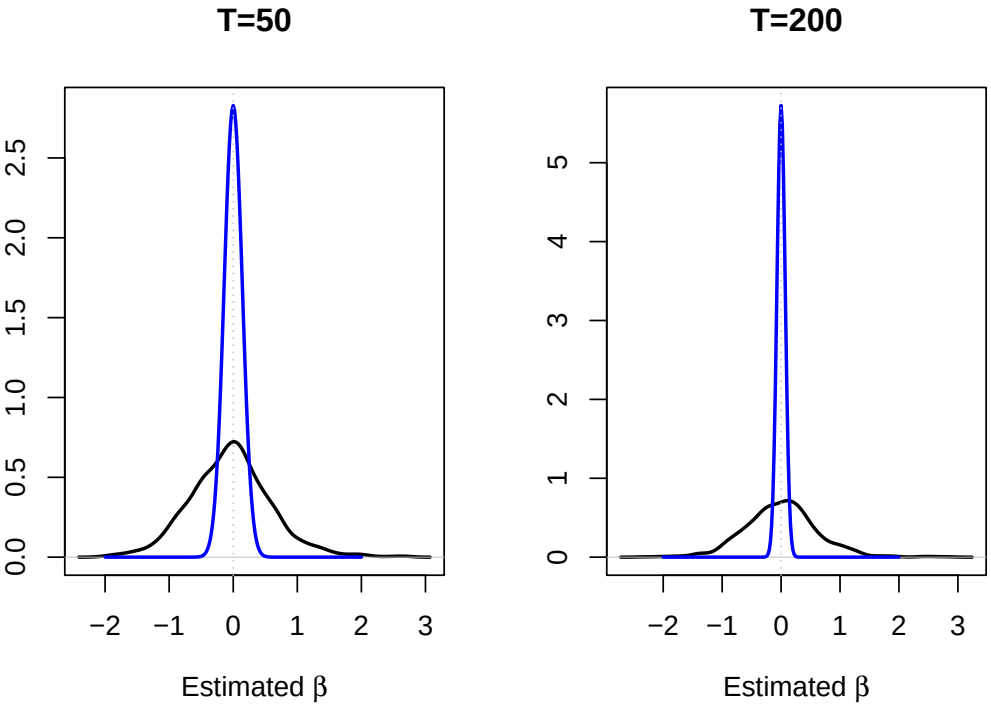


Figure 5.2: The densities of based on 1000 simulations of samples of length T, they are approximated by the kernel approach.

5.1.3 Nonstationarity tests

Hence, employing (OLS) regressions with non-stationary variables may be misleading. It is therefore important, before employing these techniques (OLS), to check that the variables are stationary. To do so, specific tests are

used. These tests are called stationarity and non-stationarity (or unit-root) tests. In the context of a stationarity test, the null hypothesis that y_t is stationary or trend-stationary (i.e. equal to the sum of a linear trend and a stationary component). In the context of a non-stationarity (or unit root) tests, the null hypothesis that y_t is not (trend) stationary.

To illustrate, consider the AR(1) case:

$$y_t = \phi y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1). \tag{5.2}$$

If y_t is stationary (i.e. if $|\phi| < 1$, Prop. 2.3), it can be shown (Hamilton (1994), p.216) that:

$$\sqrt{T}(\phi^{OLS} - \phi) \xrightarrow{d} \mathcal{N}(0, 1-\phi^2).$$

The previous equation does not make sense for $\phi = 1$.

Even if standard OLS results are not valid any more in the non-stationary case, many (non)stationary tests make use of statistics that were used in the standard OLS analysis. Even if the t-statistic does not admit the Student-t distribution any more, one can still compute this statistic. The idea behind severeal unit-root tests is simply to determine the distribution of these OLS-based statistics under the null hypothesis of non-stationarity. Let us define H_0 as follows:

$$H_0: \quad \phi = 1 \quad \textit{and} \quad H_1: \quad |\phi| < 1.$$

Test statistic:

$$t_{\phi=1} = \frac{\phi^{OLS} - 1}{\sigma_\phi^{OLS}},$$

where ϕ^{OLS} is the OLS estimate of ϕ and σ_ϕ^{OLS} is the usual OLS-based standard error estimate of ϕ^{OLS}.

Importantly, it has to be noted that, {under the null hypothesis, $t_{\phi=1}$ is not distributed as a Student-t variable} (see Def. 8.12).

Theorem 5.1 (Convergence results when $\phi = 1$)**.** *If y_t follows Eq.* (5.2)

(with $\mathbb{V}ar(\varepsilon_t) = \sigma^2$) and $\phi = 1$, then:

$$T^{-3/2} \sum_{t=1}^{T} y_t \xrightarrow{d} \sigma \int_0^1 W(r)dr \tag{5.3}$$

$$T^{-2} \sum_{t=1}^{T} y_t^2 \xrightarrow{d} \sigma^2 \int_0^1 W(r)^2 dr \tag{5.4}$$

$$T^{-1} \sum_{t=1}^{T} y_{t-1}\varepsilon_t \xrightarrow{d} \sigma^2 \int_0^1 W(r)dW(r), \tag{5.5}$$

where $W(r)$ denotes a standard Brownian motion (Wiener process) defined on the unit interval.

Proof. See Phillips (1987) or Hamilton (1994) (Subsection 17.4). □

Theorem 5.1 notably implies that, if $\phi = 1$, then:

$$T(\phi^{OLS} - 1) \xrightarrow{d} \frac{\int_0^1 W(r)dW(r)}{\int_0^1 W(r)^2 dr} = \frac{\frac{1}{2}(W(1)^2 - 1)}{\int_0^1 W(r)^2 dr} \tag{5.6}$$

$$t_{\phi=1} \xrightarrow{d} \frac{\int_0^1 W(r)dW(r)}{\left(\int_0^1 W(r)^2 dr\right)^{1/2}} = \frac{\frac{1}{2}(W(1)^2 - 1)}{\left(\int_0^1 W(r)^2 dr\right)^{1/2}} \tag{5.7}$$

The previous theorem notably implies that the convergence rate of ϕ^{OLS} (in T) is faster than if y_t was stationary (in which case it is in $\sqrt{T}$).

it also implies that, asmptotically, (a) ϕ^{OLS} is not normally distributed and (b) $t_{\phi=1}$ is not standard normal.

The limiting distribution of $t_{\phi=1}$ has no closed form; it is called the **Dickey-Fuller distribution**.

Although not available in closed form (it has to be evaluated numerically), the distribution of $T(\phi^{OLS} - 1)$ can be used to test for the null hypothesis $\phi = 1$.

Nonstationarity tests: About trends again

The right consideration for potential trends in y_t's specification is crucial.

We focus on two cases:

- Constant only. Specification:

$$y_t = c + \phi y_{t-1} + \varepsilon_t.$$

  $|\phi| < 1$: y_t is stationary (we note $I(0)$) with non-zero mean.
- Constant + Linear trend. Specification:

$$y_t = c + \delta t + \phi y_{t-1} + \varepsilon_t.$$

  $|\phi| < 1$: y_t is stationary around a deterministic time trend.

So far, we focused on AR(1) processes. But many time series have a more complicated dynamic structure.

Dickey-Fuller (DF) test

Dickey and Fuller (1979)

Specification underlying the Dickey-Fuller (DF) test:

$$y_t = \beta' D_t + \phi y_{t-1} + \sum_{i=1}^{p} \psi_i \Delta y_{t-i} + \varepsilon_t, \tag{5.8}$$

where D_t is a vector of deterministic trends ($D_t = 1$ or $D_t = [1, t]'$) and p should be such that the estimated ε_t are serially uncorrelated.

The null hypothesis of the DF test is: $\phi = 1$. Test statistics:

$$\begin{aligned} \text{ADF bias statistic:} \qquad & ADF_\pi = T(\phi^{OLS} - 1) \\ \text{ADF } t \text{ statistic:} \qquad & ADF_t = t_{\phi=1} = \frac{\phi^{OLS} - 1}{\sigma_\phi^{OLS}} \end{aligned}$$

Under the null hypothesis ($\phi = 1$) and if the regression is as in Eq. (5.2) the limiting distributions of the statistics are, respectively, as in Eq. (5.6) and Eq. (5.7).

Alternative formulation of Eq. (5.8):

$$\Delta y_t = \beta' D_t + \pi y_{t-1} + \sum_{i=1}^{p} \psi_i \Delta y_{t-i} + \varepsilon_t,$$

with $\pi = \phi - 1$. Under the null hypothesis $\pi = 0$. In this case:

$$\begin{aligned} \text{ADF bias statistic:} \qquad & ADF_\pi = T\pi^{OLS} \\ \text{ADF } t \text{ statistic:} \qquad & ADF_t = t_{\pi=0} = \frac{\pi^{OLS}}{\sigma_\pi^{OLS}} \end{aligned}$$

Under the null hypothesis ($\phi = 1$) and if the regression is as in Eq. (5.2) the limiting distributions of the statistics are, respectively, as in Eq. (5.6) and Eq. (5.7).

The selection of p can rely on information criteria (see Def. 2.11).

Alternatively, Schwert (1989) proposes to use:

$$p = \left[12 \times \left(\frac{T}{100} \right)^{1/4} \right].$$

Importantly, this test is **one-sided left-tailed** test: one rejects the null if the test statistics are sufficiently negative; we are therefore interested in the first quantiles of the limit distribution.

Phillips-Perron (PP) test

Phillips and Perron (1988)

Test regression: $\Delta y_t = \beta' D_t + \pi y_{t-1} + \varepsilon_t$.

The issue of serial correlation (and heteroskedasticity) in the residual is handled by adjusting the test statistics $t_{\pi=0}$ and $T\pi$:[1]

$$\begin{aligned} \text{PP } t \text{ stat.:} \qquad & Z_t = \sqrt{\frac{\hat{\gamma}_{0,T}}{\hat{\lambda}_T^2}} t_{\pi=0,T} - \frac{\hat{\lambda}_T^2 - \hat{\gamma}_{0,T}}{2\hat{\lambda}_T} \left(\frac{T\sigma_{\pi,T}^{OLS}}{s_T} \right) \\ \text{PP bias stat.:} \qquad & Z_\pi = T\pi_T^{OLS} - \frac{1}{2}(\hat{\lambda}^2 - \hat{\gamma}_{0,T}^2) \left(\frac{T\sigma_{\pi,T}^{OLS}}{s_T^2} \right)^2 \end{aligned}$$

[1] See Hamilton (1994), Table 17.2 p.514.

where

$$\begin{aligned}
\hat{\gamma}_{j,T} &= \frac{1}{T}\Sigma_{t=j+1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_{t-j} \\
\hat{\varepsilon}_t &= \text{OLS residuals} \\
\hat{\lambda}_T^2 &= \hat{\gamma}_{0,T} + 2\Sigma_{j=1}^{q}\left(1-\frac{j}{q+1}\right)\hat{\gamma}_{j,T} \quad \text{(Newey-West formula)} \\
s_T^2 &= \frac{1}{T-k}\Sigma_{t=1}^{T}\hat{\varepsilon}_t^2 \quad (k\text{: number of param. estim. in the regression}) \\
\sigma_{\pi,T}^{OLS} &= \text{OLS standard error of } \pi.
\end{aligned}$$

When the underlying regression is: $y_t = \alpha + \phi y_{t-1} + \varepsilon_t$, and under the null that $\alpha = 0$ and $\phi = 1$, we have that:

- the limiting distribution of Z_π is that of (see, e.g., Hamilton (1994) 17.6.8):
$$\frac{\frac{1}{2}\{W(1)^2 - 1\} - W(1)\int_0^1 W(r)dr}{\int_0^1 W(r)^2 dr - \left[\int_0^1 W(r)dr\right]^2};$$
- the limiting distribution of Z_t is that of (see, e.g., Hamilton (1994) 17.6.12):
$$\frac{\frac{1}{2}\{W(1)^2 - 1\} - W(1)\int_0^1 W(r)dr}{\left(\int_0^1 W(r)^2 dr - \left[\int_0^1 W(r)dr\right]^2\right)^2}.$$

If $|\phi| < 1$, the OLS estimate ϕ_T^{OLS} is not consistent if the ε_ts are serially correlated (the true residuals and the regressors are correlated). When $\phi = 1$, the rate of convergence of ϕ_T^{OLS} is T (*super-consistency*), which ensures that $\phi_T^{OLS} \xrightarrow{p} 1$ even if the ε_ts are serially correlated.

As the ADF test, this test is **one-sided left-tailed** (reject the null if the test statistics are sufficiently negative). The critical values are obtained by simulation; they can for instance be found here.

Stationarity test: KPSS

The test proposed by Kwiatkowski et al. (1992) is a stationarity test, i.e., under the null hypothesis, the process is stationary. The underlying specificaiton is the following:

$$y_t = \beta' D_t + \mu_t + \varepsilon_t$$

with $\mu_t = \mu_{t-1} + \eta_t$, $\mathbb{V}ar(\eta_t) = \sigma_\eta^2$, where $D_t = 1$ or $D_t = [1, t]'$ and where $\{\varepsilon_t\}$ is a covariance-stationary sequence.

The KPSS statistic corresponds to the Lagrange Multiplier test statistic associated with the hypothesis $\sigma_\eta^2 = 0$:

$$\boxed{\xi_T^{KPSS} = \left(\frac{1}{\hat{\lambda}_T^2 T^2} \sum_{t=1}^{T} \hat{S}_t^2 \right),}$$

with $\hat{S}_t = \Sigma_{i=1}^t \hat{\varepsilon}_i$, where the $\hat{\varepsilon}_t$s are the residuals of the OLS regression of y_t on D_t, and where $\hat{\lambda}^2$ is a consistent estimate of the long-run variance of $\hat{\varepsilon}_t$ (see Def. 1.9 and Newey-West approach, see Eq. (1.6)).

KPSS show that, under the null hypothesis, ξ_T^{KPSS} converges in distribution towards a distribution that does not depend on β but on the form of D_t. Specifically:

- If $D_t = 1$:

$$\xi_T^{KPSS} \xrightarrow{d} \int_0^1 (W(r) - rW(1))dr.$$

- If $D_t = [1, t]'$:

$$\xi_T^{KPSS} \xrightarrow{d} \int_0^1 \left\{ W(r) + r(2-3r)W(1) + 6r(r^2-1) \int_0^1 W(s)ds \right\} dr.$$

This test is a **one-sided right-tailed** test: one rejects the null if ξ_T^{KPSS} is above the $(1-\alpha)$ quantile of the limit distribution. The critical values car be found, e.g., here.

Example 5.1 (Stationarity of inflation and interest rates)**.** Let us use quarterly US data and test the stationarity of inflation and the 3-month short-term rate. Let us first plot the data:

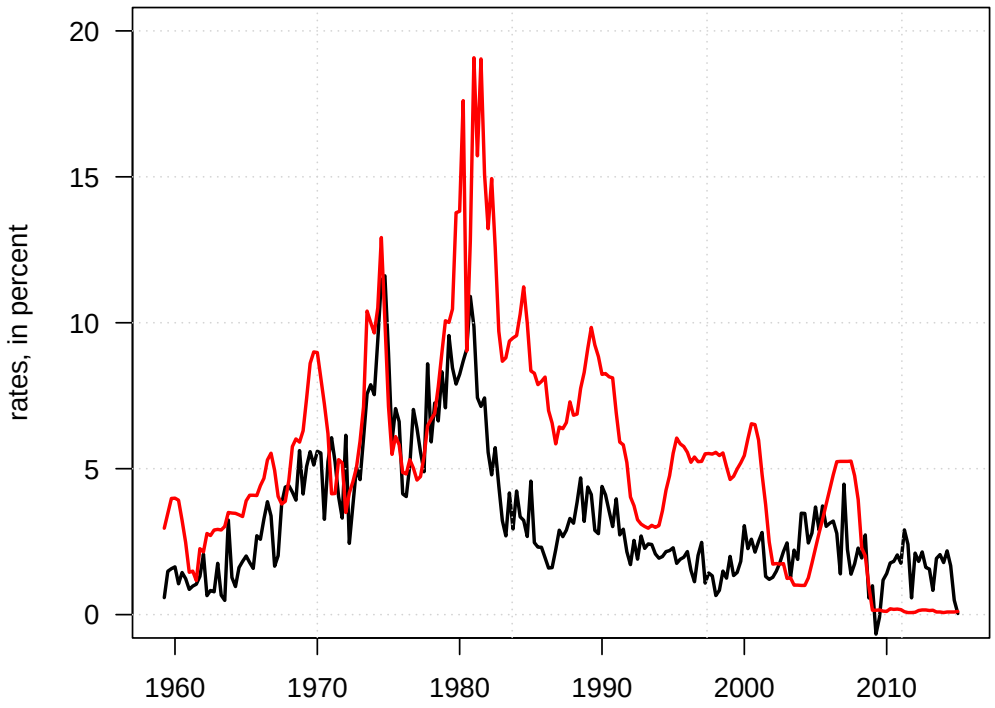


Figure 5.3: US Inflation and short-term nominal rates.

Let us now run the tests. Note that the default alternative hypothesis of function `adf.test` of package `tseries` is that the process is trend-stationary. Note that when `kpss.test` returns a p-value of `0.01`, it means that the true p-value is lower then that.

```
library(tseries)
test.adf.infl <- adf.test(US3var$infl,k=4)
test.adf.r    <- adf.test(US3var$r,k=4)
test.pp.infl <- pp.test(US3var$infl)
test.pp.r    <- pp.test(US3var$r)
test.kpss.infl <- kpss.test(US3var$infl)
test.kpss.r    <- kpss.test(US3var$r)
c(test.adf.infl$p.value,test.pp.infl$p.value,test.kpss.infl$p.value)
```

```
## [1] 0.29230878 0.04978275 0.01000000
```

```
c(test.adf.r$p.value,test.pp.r$p.value,test.kpss.r$p.value)
```

```
## [1] 0.3837577 0.3614365 0.0100000
```

Chapter 6

Introduction to cointegration

6.1 Intuition

Many statistical procedures are well-defined only when the processes of interest are stationary. As a result, especially when one wants to investigate the joint dynamics of different variables, one often begins by making the data stationary (by, e.g., taking first differences or removing deterministic trends). However, doing so may remove information from the data. Heuristically, removing trends amounts to filtering out the long-run variations of the series. However, it may be the case that the different variables interact in the short run *and* in the long run.

For instance, the left plot of Figure 6.1 suggests that the trends of x_t and y_t are positively correlated. However, the right plot shows that, for low values of h, the correlation between $x_t - x_{t-h}$ and $y_t - y_{t-h}$ is negative. This is notably the case for $h = 1$, which means that the first differences of the two variables (i.e., Δx_t and Δy_t) are negatively correlated. Hence, focusing on the first differences would lead the researcher to think that the relationship between x_t and y_t is a negative one (while it is only the case when one focuses on the high-frequency comovements between the two variables).

Definition 6.1 (Integrated variables)**.** A univariate process $\{y_t\}$ is said to be $I(d)$ if its d^{th} difference is stationary (but not its $(d-1)^{th}$ difference).

For instance:

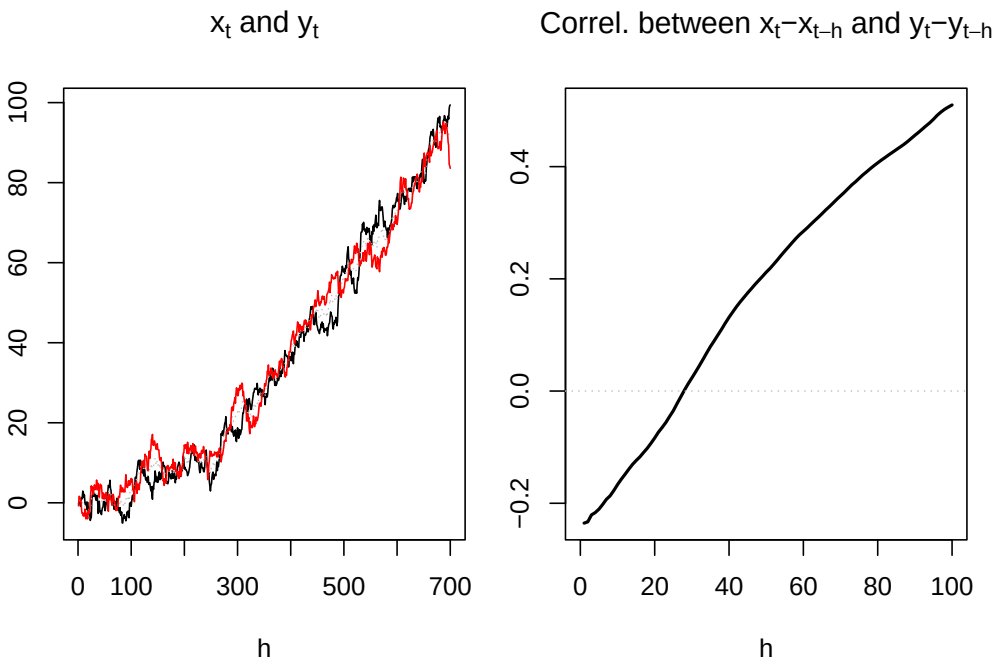


Figure 6.1: Situation where the conditional and uncondional correlation between x_t and y_t do not have the same sign.

- If y_t is not stationary but $\Delta y_t = y_t - y_{t-1}$ is, then y_t is $I(1)$.
- If Δy_t is not stationary but $\Delta^2 y_t = \Delta(\Delta y_t)$ is, then y_t is $I(2)$.
- If y_t is stationary then y_t is $I(0)$.

6.2 The bivariate case

If we regress an $I(1)$ variable y_t on another independent $I(1)$ variable x_t, the usual (OLS-based) t-tests on regression coefficients often (misleadingly) show statistically significant coefficients (we then speak of spurious regressions, see Section 5). A solution is to regress Δy_t (that is $I(0)$) on Δx_t and then inference will be correct. However, as stated above, the economic interpretation of the regression then changes, as doing so amounts to focusing on the high-frequency movements of the variables.

Let us now consider the case where y_t and x_t are still $I(1)$, but where they satisfy:

$$y_t = \beta x_t + \varepsilon_t, \tag{6.1}$$

with $\varepsilon_t \sim I(0)$. That is, there is a linear combination of the two $I(1)$ variables that is $I(0)$.

In that case, the convergence of b, the OLS estimate of β, is fast. Indeed, the convergence rate is in $1/T$ (versus $1/\sqrt{T}$ in the purely stationary case). This stems from the properties of non-statinary processes that are stated in Prop. 6.1. We have:

$$b_T = \frac{\sum_t x_t y_t}{\sum_t x_t^2} = \frac{\sum_t x_t(\beta x_t + \varepsilon_t)}{\sum_t x_t^2} = \beta + \frac{\sum_t x_t \varepsilon_t}{\sum_t x_t^2}.$$

In particular, if ε_t was a white noise, using properties (ii) and (iv) of Prop. 6.1, we would get:

$$b_T \approx \beta + \frac{1}{T} \frac{\int_0^1 x_\infty(r) dW_r^\varepsilon}{\int_0^1 x_\infty^2(r) dr}. \tag{6.2}$$

where the random variables x_∞^2 and W_r^ε are defined in Prop. 6.1.

Proposition 6.1 (Properties of an I(1) process)**.** *If $\{y_t\}$ is $I(1)$ and such that $y_t - y_{t-1} = H(L)\varepsilon_t$ where $H(L)\varepsilon_t$ is $I(0)$, then:*

i. $\frac{1}{\sqrt{T}}\bar{y}_T \xrightarrow{d} \int_0^1 y_\infty(r)dr$,
ii. $\frac{1}{T^2}\sum_{t=1}^T y_t^2 \xrightarrow{d} \int_0^1 y_\infty^2(r)dr$,
iii. $\frac{1}{T}\sum_{t=1}^T y_t(y_t - y_{t-1}) \xrightarrow{d} \frac{1}{2}y_\infty^2(1) + \frac{1}{2}\mathbb{V}ar(y_t - y_{t-1})$,
iv. $\frac{1}{T}\sum_{t=1}^T y_t\eta_t \xrightarrow{d} \sigma_\eta \int_0^1 y_\infty(r)dW_r^\eta$ *where* η_t *is a white noise of variance* σ_η^2,

where $y_\infty(r)$ *is of the form* ωW_r *(*W_r *being a Brownian motion, see* **??***), with* $\omega = \sum_{k=-\infty}^{\infty}\gamma_k$, *the* γ_k*s being the autocovariances of* $H(L)\varepsilon_t$, *and where* $\eta_\infty(r)$ *is of the form* $\sigma_\eta W_r^\eta$ *(*W_r^η *Brownian motion "associated" to* η_t*).*

Proof. (of 1) We have:

$$\frac{1}{\sqrt{T}}\bar{y}_T = \frac{1}{T}\sum_{t=1}^T \frac{y_t}{\sqrt{T}} = \frac{1}{T}\sum_{t=1}^T \tilde{y}_T(t/T),$$

where $\tilde{y}_T(r) = (1/\sqrt{T})y_{[rT]}$. Now $\tilde{y}_T(r) = r\sqrt{T}\left(\frac{1}{Tr}\sum_{t=1}^{[Tr]} H(L)\varepsilon_t\right)$. By Eq. (1.4) in Theorem 1.1, we have $\sqrt{r}\sqrt{Tr}\left(\frac{1}{Tr}\sum_{t=1}^{[Tr]} H(L)\varepsilon_t\right) \to \omega W_r$. Therefore, for large T, $\frac{1}{T}\sum_{t=1}^T \tilde{y}_T(t/T)$ approximates $\frac{1}{T}\sum_{t=1}^T \omega W_{t/T}$, which is a Riemann sum that converges to $\int_0^1 y_\infty(r)dr$. □

6.3 Multivariate case and VECM

In the following, we consider an n-dimensional vector y_t. Moreover, ε_t is an n-dimensional white noise process. The notion of integration (Def. 6.1) can also be defined in the multivariate case:

Definition 6.2 (Order of integration (multivariate case))**.** $\{y_t\}$ is $I(d)$ if

$$(1-L)^d y_t = \mu + H(L)\varepsilon_t, \tag{6.3}$$

where $H(L)\varepsilon_t$ is a stationary process (but $(1-L)^{d-1}y_t$ is not).

Definition 6.3 (Cointegration)**.** If y_t is integrated of order d, then its components are said to be cointegrated if there exists a linear combination of the components of y_t that is integrated of an order equal to, or lower than, $d-1$.

For instance, Eq. (6.1) implies that $[1, -\beta]'$ is a **cointegrating vector** and that $[x_t, y_t]'$ is cointegrated.

Consider an $I(1)$ process, y_t, that is such that the Wold representation of Δy_t is Eq. (6.3). We have:

$$y_t = \mu + H(L)\varepsilon_t + y_{t-1} = t\mu + H(L)(\varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_1) + y_0.$$

It can be shown that:

$$y_t = t\mu + H(1)(\varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_1) + \xi_t,$$

where ξ_t is a stationary process.

Assume that y_t possesses a cointegrating vector π such that $\pi' y_t$ is a (univariate) stationary process.

Necessarily, we must have $\pi'\mu = 0$ and $\pi' H(1) = 0$. Reciprocally, if $\pi'\mu = 0$ and $\pi' H(1) = 0$, then π is a cointegrating vector of y_t. This proves the following proposition:

Proposition 6.2 (Necessary and sufficient conditions of cointegration)**.** *If y_t is $I(1)$ and admits the Wold representation Eq. (6.3), with $d = 1$, then π is a cointegrating vector iff*

$$\pi'\mu = 0 \text{ (scalar equation) and } \pi' H(1) = 0 \text{ (vectorial equation).}$$

Consider the following VAR(p) model, where y_t is $I(1)$:

$$y_t = c + \Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p} + \varepsilon_t, \tag{6.4}$$

or $\Phi(L) y_t = c + \varepsilon_t$ where $\Phi(L) = I - \Phi_1 L - \cdots - \Phi_p L^p$.

Suppose that the Wold representation of Δy_t is Eq. (6.3). Premultiplying Eq. (6.3) by $\Phi(L)$ gives:

$$(1 - L)\Phi(L) y_t = \Phi(1)\mu + \Phi(L) H(L) \varepsilon_t,$$

or

$$(1-L)\varepsilon_t = \Phi(1)\mu + \Phi(L)H(L)\varepsilon_t.$$

Taking the expectation on both sides gives $\Phi(1)\mu = 0$. Therefore, for the previous equation to hold for any ε_t, we must have

$$(1-L)Id = \Phi(L)H(L).$$

The previous equality implies that $(1-z)Id = \Phi(z)H(z)$ for all z. In particular, for $z = 1$:

$$0 = \Phi(1)H(1).$$

Take any row π' of $\Phi(1)$. Since $\Phi(1)\mu = 0$, we have $\pi'\mu = 0$ (scalar equation). Since $\Phi(1)H(1) = 0$, we have $\pi'H(1) = 0$ (vectorial equation). Prop. 6.2 then implies that the rows of $\Phi(1)$ are cointegrating vectors of y_t. Therefore, if $\{a_1, ..., a_h\}$ constitutes a basis for the space of cointegrating vectors, then π can be expressed as a linear combination of the a_i's. That is, we must have:

$$\pi = [a_1 \quad a_2 \quad ... \quad a_h]b = \underbrace{A}_{n\times h}\ \underbrace{b}_{h\times 1},$$

where $A = [a_1 \quad a_2 \quad ... \quad a_h]$.

Since this is true for all rows of $\Phi(1)$, it comes that this matrix is of the form:

$$\underbrace{\Phi(1)}_{n\times n} = \underbrace{B}_{n\times h}\ \underbrace{A'}_{h\times n}. \tag{6.5}$$

This shows that the number of independent cointegrating vectors—he order of cointegration of y_t—is the rank of $\Phi(1)$. This has important implications for the dynamics of Δy_t.

Consider a process y_t whose VAR representation is as in Eq. (6.4). This VAR representation can be rewritten:

$$y_t = (c + \rho y_{t-1}) + \zeta_1 \Delta y_{t-1} + \cdots + \zeta_{p-1}\Delta y_{t-p+1} + \varepsilon_t, \tag{6.6}$$

where $\zeta_k = -\Phi_{k+1} - \cdots - \Phi_p$ and $\rho = \Phi_1 + \cdots + \Phi_p$.

Example 6.1 (VAR(2))**.** For a VAR(2) with $y_t = c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \varepsilon_t$, we have:

$$y_t = c + \{\Phi_1 + \Phi_2\}y_{t-1} + \{-\Phi_2\}\Delta y_{t-1} + \varepsilon_t,$$

Subtracting y_{t-1} from both sides of Eq. (6.6) and remarking that $-\Phi(1) = \rho - Id$ (recall that $\Phi(L) = I - \Phi_1 L - \cdots - \Phi_p L^p$), we get:

$$\Delta y_t = \{c - \underbrace{\Phi(1) y_{t-1}}_{=BA' y_{t-1}}\} + \zeta_1 \Delta y_{t-1} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + \varepsilon_t.$$

Using Eq. (6.5), and denoting by z_t the h-dimensional vector $A' y_t$, we obtain the **error correction representation** of the cointegrated variable y_t:

$$\boxed{\Delta y_t = c - B z_{t-1} + \zeta_1 \Delta y_{t-1} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + \varepsilon_t.} \tag{6.7}$$

This type of model is called **Vector Error Correction Model** (VECM). This is because the components of z_t can be considered as errors that, multiplied by the components of B generate correction forces that imply that, in the long run, the congregation relationships are satisfied. Example 6.2 illustrates this.

Example 6.2 (VECM example)**.** Consider the VAR(1) process y_t that follows:

$$y_t = \Phi_1 y_{t-1} + \varepsilon_t = \begin{bmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \end{bmatrix} y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim i.i.d. \mathcal{N}(0, \Sigma)$$

1 is an eigenvalue of Φ_1 and y_t is $I(1)$. We have:

$$\Phi(1) = Id - \begin{bmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.5 & -0.5 \\ -0.2 & 0.2 \end{bmatrix},$$

which is of rank 1. Therefore y_t is cointegrated of order 1.

For that process, Eq. (6.7) writes:

$$\Delta y_t = -\begin{bmatrix} 0.5 & -0.5 \\ -0.2 & 0.2 \end{bmatrix} y_{t-1} + \varepsilon_t = \begin{bmatrix} -0.5 \\ 0.2 \end{bmatrix} z_{t-1} + \varepsilon_t,$$

where $z_t = y_{1,t} - y_{2,t}$.

The process z_t is stationary (we have $z_t = 0.3 z_{t-1} + \varepsilon_{1,t} - \varepsilon_{2,t}$).

We say that there is a *long-run relationship* between $y_{1,t}$ and $y_{2,t}$:

When $y_{1,t}$ is substantially above $y_{2,t}$, z_t is large, the influence of $-0.5 z_t$ on $\Delta y_{1,t+1}$ is negative, which tends to "correct" $y_{1,t}$ and brings it closer to $y_{2,t}$.

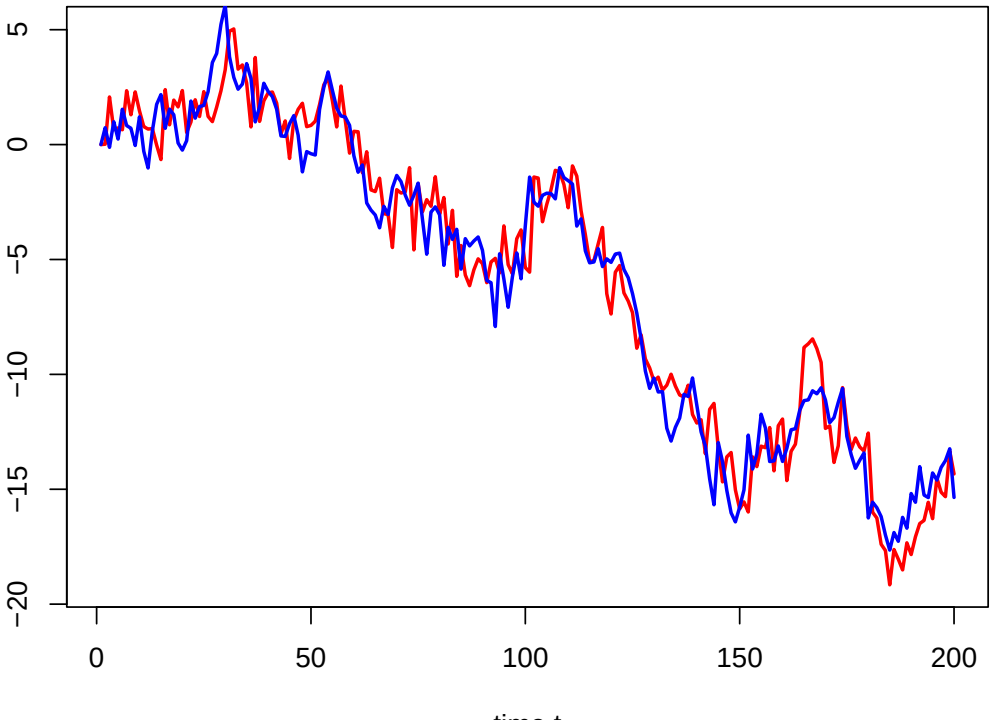


Figure 6.2: Simulation of $y_{1,t}$ (in blue) and $y_{2,t}$ (in red).

6.3.1 Cointegration in practice

Assume that we have a vector of variables y_t and that we want to investigate the joint dynamics of its components. If the $y_{i,t}$s are $I(1)$, we may need to use a VECM. Therefore, in the first place, one has to test for the stationarity of y_t's components (see Section 5).

If the components of y_t are $I(1)$, one has to determine the existence of cointegrating vectors. There are two general possibilities to do this:

1. The theory provides us with relevant cointegration relationships, for instance:
   - The Purchasing Power Parity (PPP) suggests the existence of a long-run relationship between domestic prices, foreign prices, and the exchange rate.
   - If real rates are stationary, the Fisher equation ($r = i - \pi$) implies a cointegration relatinsip between the nominal interest rate (i) and inflation π.
2. We have no a priori regarding the cointegration relationship. We have to estimate (and test) the potential cointegration relationship.

In the first case, we proceed in two steps: a. If π^* is suspected to be a cointegrating vector, we can then use unit-root or stationarity tests on $z_t = \pi^{*'} y_t$ (see Section **??**). b. Estimate Eq. (6.7) by OLS.

In the following, we focus on the second case, when there is at most one cointegration relationship. (In the general case, one can for instance implement the Johansen (1991) approach.)

Engle and Granger (1987) propose a two-step estimation procedure to estimate error-correction models. This method proceeds under the assumption that there is a single cointegration equation, i.e. $A' = [\alpha_1, ..., \alpha_n]$. (Recall that $\Phi(1) = BA'$, see Eq. (6.5).) Without loss of generality, one can set $\alpha_1 = 1$. In that case, the cointegration relationship, if it exists, is of the form:

$$y_{1,t} = -\alpha_2 y_{2,t} - \cdots - \alpha_n y_{n,t} + z_t,$$

where $z_t \sim I(0)$.

The first step consists in estimating the previous equation by OLS (regression of $y_{1,t}$ on $y_{2,t}, \dots, y_{2,t}$). The second step consists in estimating Eq. (6.7) also by OLS, after having replaced z_{t-1} by the (lagged) residuals of the first step OLS regression ($\hat{z}_{t-1}$). Because of high speed convergence of the first-step regression (the convergence is in $1/T$, see Eq. (6.2)), the asymptotic properties of the second-step estimates are the standard ones. That is, one can use the standard t-statistic to assess the statistical significativity of the parameters.

It remains to explain how to test for the existence of a cointegration relationship. This amounts to testing whether z_t is stationary. Note however that we do not observe the "true" z_t, but only OLS-based estimates $\hat{z}_t$. Therefore, the critical values of the usual unit-root tests are not the same. The appropriate critical values are given by Phillips and Ouliaris (1990).

Example 6.3 (VECM for US inflation and nominal interest rate)**.** The data are as in Example 5.1. In the first step, we compute z_t and use Phillips and Ouliaris (1990)'s test to see whether the two variables are cointegrated:

```
library(AEC);library(tseries)
T <- dim(US3var)[1]
infl <- US3var$infl
i    <- US3var$r
eq <- lm(i~infl)
z <- eq$residuals
po.test(cbind(i,infl))
```

```
##
##  Phillips-Ouliaris Cointegration Test
##
## data:  cbind(i, infl)
## Phillips-Ouliaris demeaned = -29.484, Truncation lag parameter = 2,
## p-value = 0.01
```

We reject the null hypothesis of unit root in the residuals. That is, the results are in favor of cointegration. Let us then esitmate the VECM model; this amounts to running two OLS regressions:

Table 6.1: Results of the second-stage OLS regressions (Engle-Granger approach). The first variable is the short-term nominal interest rate; the second variable is inflation.

	Coef Y1	p-values Y1	Coef Y2	p-values Y2
(Intercept)	-0.0189	0.8093	-0.0047	0.9457
z_1	-0.1040	0.0014	-0.0034	0.9046
dinfl_1	0.0224	0.7852	-0.3845	0.0000
di_1	-0.0841	0.2225	0.1196	0.0473
dinfl_2	0.0322	0.6864	-0.2238	0.0014
di_2	-0.0393	0.5625	0.0828	0.1626

```
infl_1 <- c(NaN,infl[1:(T-1)])
infl_2 <- c(NaN,NaN,infl[1:(T-2)])
infl_3 <- c(NaN,NaN,NaN,infl[1:(T-3)])
i_1 <- c(NaN,i[1:(T-1)])
i_2 <- c(NaN,NaN,i[1:(T-2)])
i_3 <- c(NaN,NaN,NaN,i[1:(T-3)])
z_1 <- c(NaN,z[1:(T-1)])
dinfl   <- infl   - infl_1
dinfl_1 <- infl_1 - infl_2
dinfl_2 <- infl_2 - infl_3
di   <- i   - i_1
di_1 <- i_1 - i_2
di_2 <- i_2 - i_3
eq1 <- lm(di    ~ z_1 + dinfl_1 + di_1 + dinfl_2 + di_2)
eq2 <- lm(dinfl ~ z_1 + dinfl_1 + di_1 + dinfl_2 + di_2)
```

Table 6.1 reports the estimated coefficients in `eq1` and `eq2`, together with their p-values:

Imporantly, one of the parameters associated with z_{t-1} (parameters that are called speeds of adjustement) is significant and with the expected signs: if the short term nominal interest rate i_t is high then z_t becomes positive (since, up to an intercept, $z_t = i_t - \beta\pi_t$, where β comes from the first-step regression), but since $B_{1,1}$ is negative (it is equal to -0.104), a positive z_t will generate a negative correction (i.e., $B_{1,1}z_t$) for i_{t+1} on date $t+1$, thereby "correcting"

the high level of i_t.

Chapter 7

ARCH and GARCH Models

7.1 Conditional heteroskedasticity

Many financial and macroeconomic variables are hit by shocks whose variance is not constant through time, i.e. by *heteroskedatic* shocks. Often, the conditional variance of shocks features a persistent behavior (volatility clustering). The observation that large shocks (in absolute value) tend to be followed by other important shocks has notably been established by Mandelbrot (1963). Such a situation is illustrated by Figure 7.1.

Autoregressive Conditional Heteroskedasticity (ARCH) and its generalized version (GARCH) constitute useful tools to model such time series.

In order to test whether a times series of shocks $\{u_1, ..., u_T\}$ features persistent conditional heteroskedasticity, a simple test has been designed by Engle (1982). This test consists in regressing u_t^2 on it last m lags (i.e., on $u_{t-1}^2, ..., u_{t-m}^2$) by OLS.Under the null hypothesis that u_t is i.i.d., we have:

$$T \times R^2 \xrightarrow{d} \chi^2(m),$$

where R^2 is the centered R^2 of the OLS regression.

Let us employ this test on the return series plotted in the lower panel of Figure 7.1:

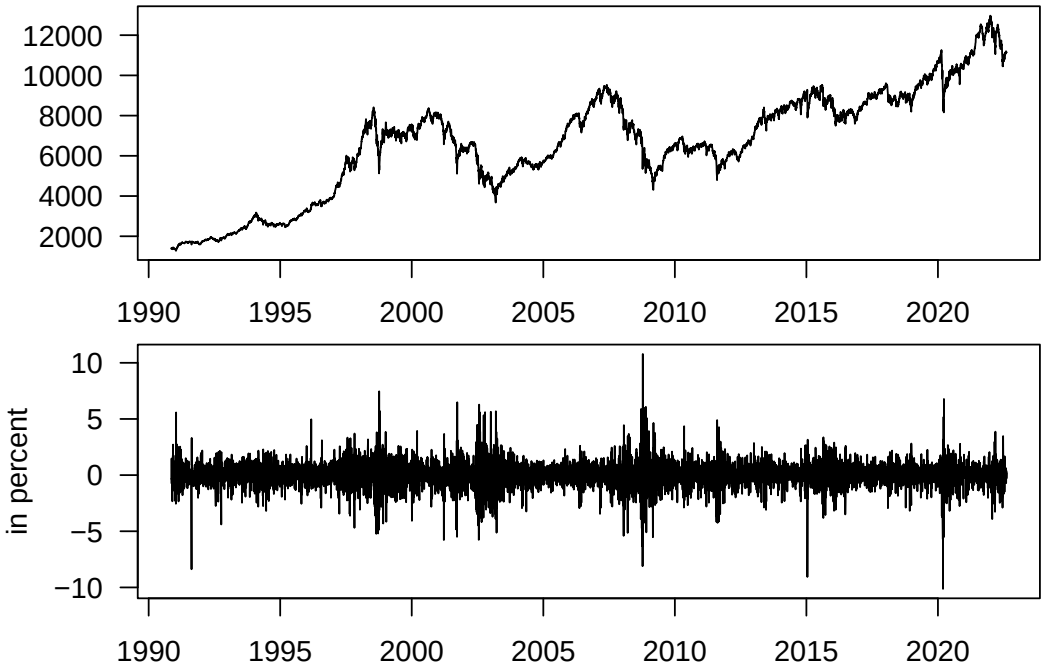


Figure 7.1: Upper plot: SMI index (daily Close prices); lower plot: daily log returns.

```
library(AEC);data(smi)
smi <- smi[complete.cases(smi),] # remove NaNs
T <- dim(smi)[1]
smi$r <- 100*c(NaN,log(smi$Close[2:T]/smi$Close[1:(T-1)]))
u <- smi$r^2
u_1 <- c(NaN,smi$r[1:(T-1)]^2)
u_2 <- c(NaN,NaN,smi$r[1:(T-2)]^2)
eq <- lm(u^2 ~ u_1^2 + u_2^2)
test.stat <- length(u)*summary(eq)$r.squared
pvalue <- 1 - pchisq(q = test.stat,df=2)
```

The p-value is extremely close to zero. Hence we strongly reject the null of an i.i.d. SMI return.

7.2 The ARCH model

7.2.1 The two ARCH specifications

Consider the auto-regressive process following:

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + u_t, \tag{7.1}$$

where u_t is a white noise (Def. 1.1), i.e. $\mathbb{E}(u_t) = 0$, $\mathbb{E}(u_t^2) = \sigma^2$ and $\mathbb{E}(u_t u_s) = 0$ if $s \neq t$. Importantly, note that while the *unconditional variance* of u_t is σ^2, its *conditional variance* can be time-varying. This is the case in the context of (G)ARCH models.

In the ARCH(m) model, u_t follows:

$$u_t^2 = \zeta + \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2 + w_t, \tag{7.2}$$

where w_t is a non-autocorrelated white noise process that is exogenous to u_t, in the sense that:

$$\mathbb{E}(w_t | u_{t-1}, u_{t-2}, \dots) = 0.$$

In this case:

$$\mathbb{E}(u_t^2 | u_{t-1}, u_{t-2}, \dots) = \zeta + \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2. \tag{7.3}$$

For Eq. (7.2) to make sense, it has to be the case that

$$\zeta + \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2 + w_t \leq 0$$

for all realisations of $\{u_t\}$. This is the case if $w_t > -\zeta$, with $\zeta > 0$, and if $\alpha_i \geq 0$ for $i \in \{1, \ldots, m\}$. Assuming this is the case, u_t^2 is covariance-stationary if the roots of:

$$g(z) = 1 - \alpha_1 z - \cdots - \alpha_m z^m = 0$$

lie outside the unit circle (Prop. **??**). A necessary condition for not having a root between 0 and 1 is that $\sum_i \alpha_i < 1$.[1] This is also a sufficient condition.[2]

If $w_t > -\zeta$, with $\zeta > 0$, $\alpha_i \geq 0$ for $i \in \{1, \ldots, m\}$ and $\sum_i \alpha_i < 1$, then the unconditional variance of u_t is:

$$\sigma^2 = \mathbb{E}(u_t^2) = \frac{\zeta}{1 - \sum_{i=1}^m \alpha_i}.$$

An alternative representation of an ARCH(m) process is as follows:

$$u_t = \sqrt{h_t} v_t, \tag{7.4}$$

where v_t is an i.i.d. sequence with zero mean and unit variance, i.e.:

$$\mathbb{E}(v_t) = 0, \quad \mathbb{E}(v_t^2) = 1.$$

If h_t follows:

$$h_t = \zeta + \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2, \tag{7.5}$$

then Eq. (7.3) is also true. Since $u_t^2 = h_t v_t^2$, Eq. (7.2) holds with $w_t = h_t(v_t^2 - 1)$. This alternative representation is convenient because v_t is not necessarily bounded.

[1] If we have $1 - \sum_i \alpha_i \leq 0$, then $g(1) \leq 0$. We would then have $g(0) = 1 > 0$ and $g(1) \leq 0$. Since g is continuous, this would imply that $\exists z \in [0, 1]$ s.t. $g(z) = 0$.

[2] Indeed, for $z \in [-1, 1]$, we have $1 - \alpha_1 z - \cdots - \alpha_m z^m \geq 1 - \sum_i |\alpha_i z^i| \geq 1 - \sum_i \alpha_i > 0$.

7.2.2 Maximum Likelihood Estimation of an ARCH process

Assume the complete model is:

$$y_t = \mathbf{x}_t'\beta + u_t,$$

where $\mathbf{x}_t$ is a $k \times 1$ vector of explanatory variables and u_t is as specified in Eq. (7.4).

To write the likelihood function, it is convenient to condition on the first m observations. Let us denote by $\mathcal{I}_t$ the following information set:

$$\mathcal{I}_t = (y_t, y_{t-1}, \ldots, y_0, y_{-1}, \ldots, y_{-m+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_0, \mathbf{x}_{-1}, \ldots, \mathbf{x}_{-m+1}).$$

If $v_t \sim \mathcal{N}(0,1)$, we have, for $t \geq 1$:

$$f(y_t|\mathbf{x_t}, \mathcal{I}_{t-1}) = \frac{1}{\sqrt{2\pi h_t}} \exp\left(\frac{-(y_t - \mathbf{x}_t'\beta)^2}{2h_t}\right), \tag{7.6}$$

where h_t is given by:

$$h_t = \zeta + \alpha_1(y_{t-1} - \mathbf{x}_{t-1}'\beta)^2 + \cdots + \alpha_m(y_{t-m} - \mathbf{x}_{t-m}'\beta)^2.$$

The log likelihood function is given by:

$$\log \mathcal{L}(\theta) = \sum_{t=1}^{T} \log(f(y_t|\mathbf{x_t}, \mathcal{I}_{t-1})),$$

where θ, the vector of unknown parameters, is $[\beta', \zeta, \alpha']'$, where $\alpha = [\alpha_1, \ldots, \alpha_m]'$. The maximisation of the log likelihood is performed numerically.

Note that one can also use non-Gaussian distributions for v_t. (For that, one has to replace the normal distribution in Eq. (7.6).)

Let us fit an ARCH(2) model on the SMI return data (lower plot of Figure 7.1). For this, we make use of function `compute.garch` of package `AEC`. This function takes four arguments:

- vector `theta` contains the model parameterization: first, ζ, then the α_i's, then for GARCH models (see Subsection 7.3 below), the δ_i's.
- vector `x` contains the observations of the process.
- `m` is the number of lags in the ARCH specification.
- `r` is the number of lags in the GARCH specification (see Subsection 7.3 below).

Function `compute.garch` returns a list, one entry of each is the log-likelihood associated with a parameterization $[\zeta, \alpha']'$, the second being the resulting sequence of h_t's (see Eq. (7.5)). Let us create a function that returns only the log-likelihood:

```
loglik <- function(theta,x,m,r){
  # first parameter of theta: zeta
  # next: alpha's (ARCH)
  # next: delta's (GARCH)
  Garch <- compute.garch(theta,x,m,r)
  return(Garch$logf)
}
```

Now, let us maximize the log-likelihood:

```
m <- 2
r <- 0 # for ARCH models, r=0
smi <- smi[4000:dim(smi)[1],] # reduce sample
par0 <- c(0.62,0.2,0.2)
res.opt <- optim(par=par0,x=smi$r,m=m,r=r,loglik,
                 method="BFGS",hessian=TRUE,
                 control = list(trace=TRUE,maxit = 10))
estim.param <- res.opt$par
std.dev <- sqrt(diag(solve(res.opt$hessian)))
t.stat <- estim.param/std.dev
```

Table 7.1 reports the estimated parameters and their standard deviation. Figure 7.2 displays the resulting 95% confidence intervals (i.e., $\pm 2\sqrt{h_t}$).

Table 7.1: ARCH(2), ML estimation results. Data: SMI daily returns.

	Estim. param.	Std dev.	t-stat
zeta	0.6042	0.0230	26.3054
alpha1	0.2939	0.0289	10.1704
alpha2	0.2321	0.0250	9.2838

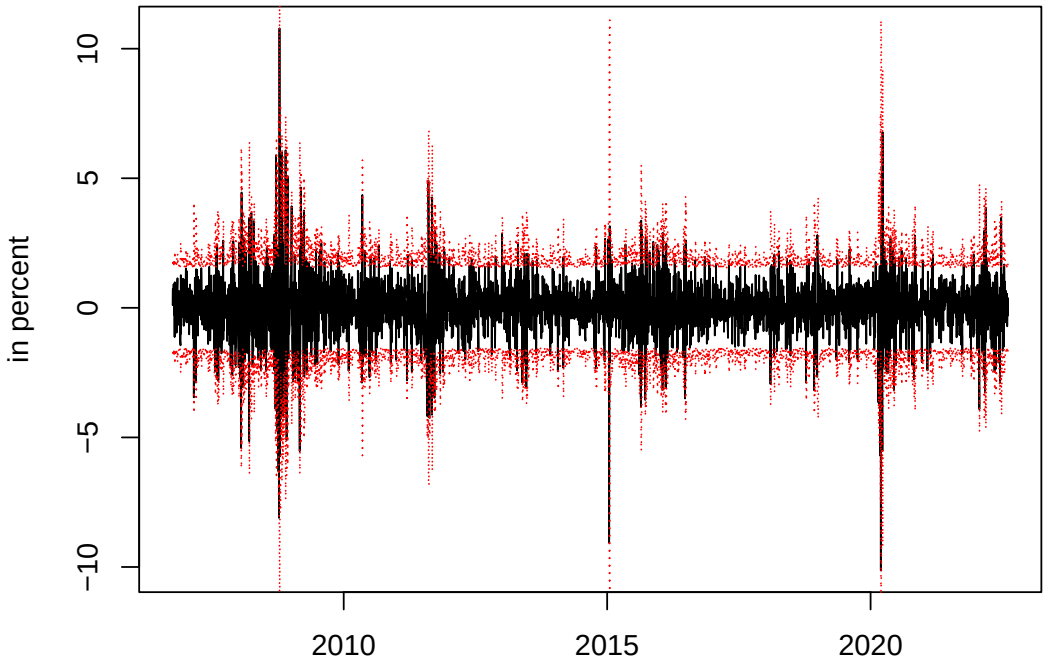


Figure 7.2: SMI daily returns (in black) and, in red, 95% confidence interval based on the ARCH(2) estimated model ($\pm 2\sqrt{h_t}$).

7.3 The GARCH model

One can generalize the model and replace Eq. (7.5) with:

$$\begin{aligned} h_t &= (1-\delta_1-\delta_2-\cdots-\delta_r)\zeta + \\ &\quad \delta_1 h_{t-1} + \delta_2 h_{t-2} + \cdots + \delta_r h_{t-r} + \\ &\quad \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2. \end{aligned} \tag{7.7}$$

This generalised autoregressive conditional heteroskedasticity model is denoted by GARCH(r,m). Non-negativity is satisfied as soon as $\kappa > 0$, $\alpha_j \geq 0$, $\delta_j \geq 0$ for $j \leq p$.

Denoting $u_t^2 - h_t$ by w_t, and $(1-\delta_1-\cdots-\delta_r)\zeta$ by κ, it can be checked that:[3]

$$\begin{aligned} u_t^2 &= \kappa + (\delta_1+\alpha_1)u_{t-1}^2 + (\delta_2+\alpha_2)u_{t-2}^2 + ... \\ &\quad +(\delta_p+\alpha_p)u_{t-p}^2 + w_t - \delta_1 w_{t-1} - \cdots - \delta_r w_{t-r}, \end{aligned} \tag{7.8}$$

where $p = max(m,r)$, $\alpha_j = 0$ for $j > m$ and $\delta_j = 0$ for $j > r$.

We have $w_t = h_t(v_t^2 - 1)$. Under regularity assumptions, $\{w_t\}$ is a white noise. Hence, u_t^2 follows an ARMA(p,r) process. Accordingly, it comes that u_t^2 is covariance stationary if the roots of:

$$1 - (\delta_1+\alpha_1)z - \cdots - (\delta_1+\alpha_1)z^p = 0$$

lie outside the unit circle (Prop. **??**). When the $\delta_i + \alpha_i$ are nonnegative, and using the same reasoning as for ARCH models, this is the case iff:

$$\sum_i \delta_i + \sum_i \alpha_i < 1.$$

In that case, the unconditional variance of u_t, i.e. the unconditional mean of u_t^2, is:

$$\mathbb{E}(u_t^2) = \sigma^2 = \frac{\kappa}{1 - \sum_i \delta_i - \sum_i \alpha_i}.$$

GARCH models can also be estimated by the ML approach.Table 7.2 reports the estimated parameters when fitting an GARCH(1,1) model on the SMI return dataset. Figure

Figure 7.3 compare the condiitonal standard deviations ($\sqrt{h_t}$) resulting from the ARCH(2) and the GARCH(1,1) specifications.

[3]Note that $w_t = h_t(v_t^2 - 1)$ is a martingale difference sequence (see Def. 1.2) because v_t is a zero-mean unit-root i.i.d. sequence.

Table 7.2: ARCH(2), ML estimation results. Data: SMI daily returns.

	Estim. param.	Std dev.	t-stat
zeta	0.2195	0.0216	10.1753
alpha1	0.1450	0.0152	9.5297
delta1	0.8299	0.0152	54.5175

```
## initial  value 5703.217726
## iter  10 value 5359.604659
## final  value 5359.604659
## stopped after 10 iterations
```

7.3.1 ARCH-in-mean

Another extension of the ARCH model is the **ARCH-in-Mean**, or ARCH-M model. That model is close to that specified by Eqs. (7.1), (7.4) and (7.5), but it also allows for a potential effect of h_t on $\mathbb{E}_{t-1}(y_t)$:

$$\begin{aligned} y_t &= c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \delta h_t + u_t \\ u_t &= \sqrt{h_t} v_t \\ h_t &= \zeta + \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2, \end{aligned}$$

where v_t is a zero-mean unit-variance i.i.d. sequence.

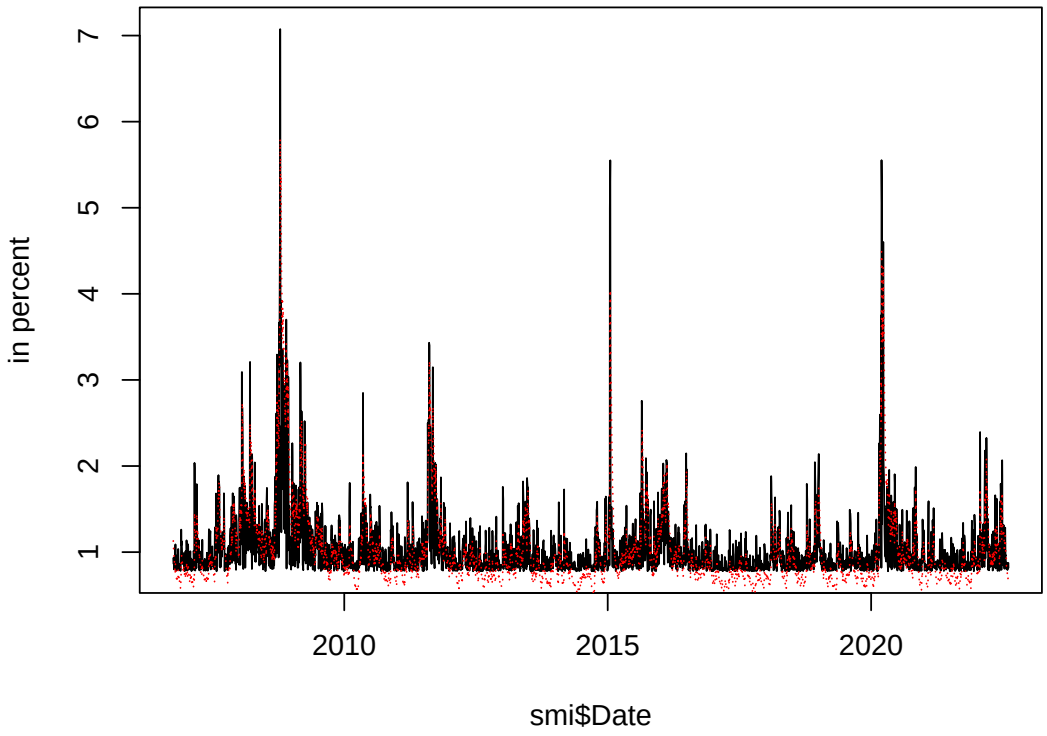


Figure 7.3: Estimated conditional standard deviations ($\sqrt{h_t}$) of SMI daily returns: ARCH(2) [black line] and GARCH(11) [dotted red line] specifications.

Chapter 8

Appendix

8.1 Principal component analysis (PCA)

Principal component analysis (PCA) is a classical and easy-to-use statistical method to reduce the dimension of large datasets containing variables that are linearly driven by a relatively small number of factors. This approach is widely used in data analysis and image compression.

Suppose that we have T observations of a n-dimensional random vector x, denoted by $x_1, x_2, \dots, x_T$. We suppose that each component of x is of mean zero.

Let denote with X the matrix given by $\begin{bmatrix} x_1 & x_2 & \dots & x_T \end{bmatrix}'$. Denote the j^{th} column of X by X_j.

We want to find the linear combination of the x_i's $(x.u)$, with $\|u\| = 1$, with "maximum variance." That is, we want to solve:

$$\begin{array}{ll} \arg\max\limits_{u} & u'X'Xu. \\ \text{s.t.} & |u| = 1 \end{array} \tag{8.1}$$

Since $X'X$ is a positive definite matrix, it admits the following decomposi-

tion:

$$\begin{aligned} X'X &= PDP' \\ &= P \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} P', \end{aligned}$$

where P is an orthogonal matrix whose columns are the eigenvectors of $X'X$.

We can order the eigenvalues such that $\lambda_1 \geq ... \geq \lambda_n$. (Since $X'X$ is positive definite, all these eigenvalues are positive.)

Since P is orthogonal, we have $u'X'Xu = u'PDP'u = y'Dy$ where $\|y\| = 1$. Therefore, we have $y_i^2 \leq 1$ for any $i \leq n$.

As a consequence:

$$y'Dy = \sum_{i=1}^{n} y_i^2 \lambda_i \leq \lambda_1 \sum_{i=1}^{n} y_i^2 = \lambda_1.$$

It is easily seen that the maximum is reached for $y = [1, 0, \cdots, 0]'$. Therefore, the maximum of the optimization program (Eq. (8.1)) is obtained for $u = P[1, 0, \cdots, 0]'$. That is, u is the eigenvector of $X'X$ that is associated with its larger eigenvalue (first column of P).

Let us denote with F the vector that is given by the matrix product XP (note that its last column is equal to Xu). The columns of F, denoted by F_j, are called **factors**. We have:

$$F'F = P'X'XP = D.$$

Therefore, in particular, the F_j's are orthogonal.

Since $X = FP'$, the X_j's are linear combinations of the factors. Let us then denote with $\hat{X}_{i,j}$ the part of X_i that is explained by factor F_j, we have:

$$\begin{aligned} \hat{X}_{i,j} &= p_{ij} F_j \\ X_i &= \sum_j \hat{X}_{i,j} = \sum_j p_{ij} F_j. \end{aligned}$$

Consider the share of variance that is explained –through the n variables $(X_1, ..., X_n)$– by the first factor F_1:

$$\frac{\sum_i \hat{X}_{i,1}\hat{X}'_{i,1}}{\sum_i X_i X'_i} = \frac{\sum_i p_{i1}F_1F'_1p_{i1}}{tr(X'X)} = \frac{\sum_i p_{i1}^2\lambda_1}{tr(X'X)} = \frac{\lambda_1}{\sum_i \lambda_i}.$$

Intuitively, if the first eigenvalue is large, it means that the first factor embed a large share of the fluctutaions of the n X_i's.

Let us illustrate PCA on the term structure of yields. The term strucutre of yields (or yield curve) is know to be driven by only a small number of factors (e.g., Litterman and Scheinkman (1991)). One can typically employ PCA to recover such factors. The data used in the example below are taken from the Fred database (tickers: "DGS6MO","DGS1", ...). The second plot shows the factor loardings, that indicate that the first factor is a level factor (loadings = black line), the second factor is a slope factor (loadings = blue line), the third factor is a curvature factor (loadings = red line).

To run a PCA, one simply has to apply function `prcomp` to a matrix of data:

```
library(AEC)
USyields <- USyields[complete.cases(USyields),]
yds <- USyields[c("Y1","Y2","Y3","Y5","Y7","Y10","Y20","Y30")]
PCA.yds <- prcomp(yds,center=TRUE,scale. = TRUE)
```

Let us know visualize some results. The first plot of Figure 8.1 shows the share of total variance explained by the different principal components (PCs). The second plot shows the facotr loadings. The two bottom plots show how yields (in black) are fitted by linear combinations of the first two PCs only.

```
par(mfrow=c(2,2))
par(plt=c(.1,.95,.2,.8))
barplot(PCA.yds$sdev^2/sum(PCA.yds$sdev^2),
        main="Share of variance expl. by PC's")
axis(1, at=1:dim(yds)[2], labels=colnames(PCA.yds$x))
nb.PC <- 2
plot(-PCA.yds$rotation[,1],type="l",lwd=2,ylim=c(-1,1),
     main="Factor loadings (1st 3 PCs)",xaxt="n",xlab="")
```

```
axis(1, at=1:dim(yds)[2], labels=colnames(yds))
lines(PCA.yds$rotation[,2],type="l",lwd=2,col="blue")
lines(PCA.yds$rotation[,3],type="l",lwd=2,col="red")
Y1.hat <- PCA.yds$x[,1:nb.PC] %*% PCA.yds$rotation["Y1",1:2]
Y1.hat <- mean(USyields$Y1) + sd(USyields$Y1) * Y1.hat
plot(USyields$date,USyields$Y1,type="l",lwd=2,
     main="Fit of 1-year yields (2 PCs)",
     ylab="Obs (black) / Fitted by 2PCs (dashed blue)")
lines(USyields$date,Y1.hat,col="blue",lty=2,lwd=2)
Y10.hat <- PCA.yds$x[,1:nb.PC] %*% PCA.yds$rotation["Y10",1:2]
Y10.hat <- mean(USyields$Y10) + sd(USyields$Y10) * Y10.hat
plot(USyields$date,USyields$Y10,type="l",lwd=2,
     main="Fit of 10-year yields (2 PCs)",
     ylab="Obs (black) / Fitted by 2PCs (dashed blue)")
lines(USyields$date,Y10.hat,col="blue",lty=2,lwd=2)
```

8.2 Linear algebra: definitions and results

Definition 8.1 (Eigenvalues)**.** The eigenvalues of of a matrix M are the numbers λ for which:

$$|M - \lambda I| = 0,$$

where $|\bullet|$ is the determinant operator.

Proposition 8.1 (Properties of the determinant)**.** *We have:*

- $|MN| = |M| \times |N|$.
- $|M^{-1}| = |M|^{-1}$.
- *If M admits the diagonal representation $M = TDT^{-1}$, where D is a diagonal matrix whose diagonal entries are $\{\lambda_i\}_{i=1,\dots,n}$, then:*

$$|M - \lambda I| = \prod_{i=1}^{n} (\lambda_i - \lambda).$$

Definition 8.2 (Moore-Penrose inverse)**.** If $M \in \mathbb{R}^{m \times n}$, then its Moore-Penrose pseudo inverse (exists and) is the unique matrix $M^* \in \mathbb{R}^{n \times m}$ that satisfies:

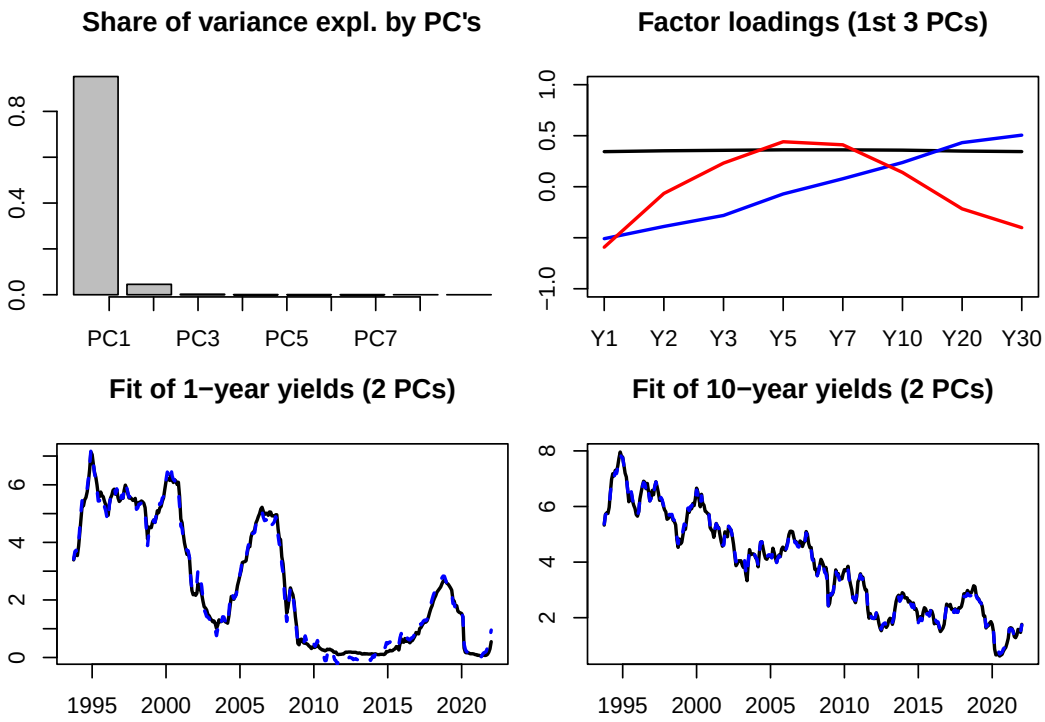


Figure 8.1: Some PCA results. The dataset contains 8 time series of U.S. interest rates of different maturities.

i. $MM^*M = M$
ii. $M^*MM^* = M^*$
iii. $(MM^*)' = MM^*$ .iv $(M^*M)' = M^*M$.

Proposition 8.2 (Properties of the Moore-Penrose inverse)**.**

- *If M is invertible then $M^* = M^{-1}$.*
- *The pseudo-inverse of a zero matrix is its transpose. **
- *
  The pseudo-inverse of the pseudo-inverse is the original matrix.

Definition 8.3 (Idempotent matrix)**.** Matrix M is idempotent if $M^2 = M$.

If M is a symmetric idempotent matrix, then $M'M = M$.

Proposition 8.3 (Roots of an idempotent matrix)**.** *The eigenvalues of an idempotent matrix are either 1 or 0.*

Proof. If λ is an eigenvalue of an idempotent matrix M then $\exists x \neq 0$ s.t. $Mx = \lambda x$. Hence $M^2x = \lambda Mx \Rightarrow (1-\lambda)Mx = 0$. Either all element of Mx are zero, in which case $\lambda = 0$ or at least one element of Mx is nonzero, in which case $\lambda = 1$. □

Proposition 8.4 (Idempotent matrix and chi-square distribution)**.** *The rank of a symmetric idempotent matrix is equal to its trace.*

Proof. The result follows from Prop. 8.3, combined with the fact that the rank of a symmetric matrix is equal to the number of its nonzero eigenvalues. □

Proposition 8.5 (Constrained least squares)**.** *The solution of the following optimisation problem:*

$$\min_{\beta} \quad ||\mathbf{y} - \mathbf{X}\beta||^2$$
$$\text{subject to } \mathbf{R}\beta = \mathbf{q}$$

is given by:

$$\boxed{\beta^r = \beta_0 - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{R}\beta_0 - \mathbf{q}),}$$

where $\beta_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Proof. See for instance Jackman, 2007. □

Proposition 8.6 (Inverse of a partitioned matrix)**.** *We have:*

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \\ -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{bmatrix}.$$

Definition 8.4 (Matrix derivatives)**.** Consider a fonction $f : \mathbb{R}^K \to \mathbb{R}$. Its first-order derivative is:

$$\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = \begin{bmatrix} \frac{\partial f}{\partial b_1}(\mathbf{b}) \\ \vdots \\ \frac{\partial f}{\partial b_K}(\mathbf{b}) \end{bmatrix}.$$

We use the notation:

$$\frac{\partial f}{\partial \mathbf{b}'}(\mathbf{b}) = \left(\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b})\right)'.$$

Proposition 8.7. *We have:*

- *If* $f(\mathbf{b}) = A'\mathbf{b}$ *where* A *is a* $K \times 1$ *vector then* $\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = A$.
- *If* $f(\mathbf{b}) = \mathbf{b}'A\mathbf{b}$ *where* A *is a* $K \times K$ *matrix, then* $\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = 2A\mathbf{b}$.

Proposition 8.8 (Square and absolute summability)**.** *We have:*

$$\underbrace{\sum_{i=0}^{\infty} |\theta_i| < +\infty}_{\textit{Absolute summability}} \quad \Rightarrow \quad \underbrace{\sum_{i=0}^{\infty} \theta_i^2 < +\infty}_{\textit{Square summability}}.$$

Proof. See Appendix 3.A in Hamilton. Idea: Absolute summability implies that there exist N such that, for $j > N$, $|\theta_j| < 1$ (deduced from Cauchy criterion, Theorem 8.2 and therefore $\theta_j^2 < |\theta_j|$. □

8.3 Statistical analysis: definitions and results

8.3.1 Moments and statistics

Definition 8.5 (Partial correlation)**.** The **partial correlation** between y and z, controlling for some variables $\mathbf{X}$ is the sample correlation between y^* and z^*, where the latter two variables are the residuals in regressions of y on $\mathbf{X}$ and of z on $\mathbf{X}$, respectively.

This correlation is denoted by $r_{yz}^{\mathbf{X}}$. By definition, we have:

$$r_{yz}^{\mathbf{X}} = \frac{\mathbf{z}^{*\prime}\mathbf{y}^*}{\sqrt{(\mathbf{z}^{*\prime}\mathbf{z}^*)(\mathbf{y}^{*\prime}\mathbf{y}^*)}}. \tag{8.2}$$

Definition 8.6 (Skewness and kurtosis)**.** Let Y be a random variable whose fourth moment exists. The expectation of Y is denoted by μ.

- The skewness of Y is given by:

$$\frac{\mathbb{E}[(Y-\mu)^3]}{\{\mathbb{E}[(Y-\mu)^2]\}^{3/2}}.$$

- The kurtosis of Y is given by:

$$\frac{\mathbb{E}[(Y-\mu)^4]}{\{\mathbb{E}[(Y-\mu)^2]\}^2}.$$

Theorem 8.1 (Cauchy-Schwarz inequality)**.** *We have:*

$$|\mathbb{C}ov(X,Y)| \leq \sqrt{\mathbb{V}ar(X)\mathbb{V}ar(Y)}$$

and, if $X \neq=$ and $Y \neq 0$, the equality holds iff X and Y are the same up to an affine transformation.

Proof. If $\mathbb{V}ar(X) = 0$, this is trivial. If this is not the case, then let's define Z as $Z = Y - \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$. It is easily seen that $\mathbb{C}ov(X,Z) = 0$. Then, the

variance of $Y = Z + \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$ is equal to the sum of the variance of Z and of the variance of $\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$, that is:

$$\mathbb{V}ar(Y) = \mathbb{V}ar(Z) + \left(\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}\right)^2 \mathbb{V}ar(X) \geq \left(\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}\right)^2 \mathbb{V}ar(X).$$

The equality holds iff $\mathbb{V}ar(Z) = 0$, i.e. iff $Y = \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X + cst$. □

Definition 8.7 (Asymptotic level)**.** An asymptotic test with critical region Ω_n has an asymptotic level equal to α if:

$$\sup_{\theta\in\Theta} \lim_{n\to\infty} \mathbb{P}_\theta(S_n \in \Omega_n) = \alpha,$$

where S_n is the test statistic and Θ is such that the null hypothesis H_0 is equivalent to $\theta \in \Theta$.

Definition 8.8 (Asymptotically consistent test)**.** An asymptotic test with critical region Ω_n is consistent if:

$$\forall \theta \in \Theta^c, \quad \mathbb{P}_\theta(S_n \in \Omega_n) \to 1,$$

where S_n is the test statistic and Θ^c is such that the null hypothesis H_0 is equivalent to $\theta \notin \Theta^c$.

Definition 8.9 (Kullback discrepancy)**.** Given two p.d.f. f and f^*, the Kullback discrepancy is defined by:

$$I(f, f^*) = \mathbb{E}^*\left(\log \frac{f^*(Y)}{f(Y)}\right) = \int \log \frac{f^*(y)}{f(y)} f^*(y)dy.$$

Proposition 8.9 (Properties of the Kullback discrepancy)**.** *We have:*

i. $I(f, f^*) \geq 0$
ii. $I(f, f^*) = 0$ *iff* $f \equiv f^*$.

Proof. $x \to -\log(x)$ is a convex function. Therefore $\mathbb{E}^*(-\log f(Y)/f^*(Y)) \geq -\log \mathbb{E}^*(f(Y)/f^*(Y)) = 0$ (proves (i)). Since $x \to -\log(x)$ is strictly convex, equality in (i) holds if and only if $f(Y)/f^*(Y)$ is constant (proves (ii)). □

Definition 8.10 (Characteristic function)**.** For any real-valued random variable X, the characteristic function is defined by:

$$\phi_X : u \to \mathbb{E}[\exp(iuX)].$$

8.3.2 Standard distributions

Definition 8.11 (F distribution)**.** Consider $n = n_1 + n_2$ i.i.d. $\mathcal{N}(0,1)$ r.v. X_i. If the r.v. F is defined by:

$$F = \frac{\sum_{i=1}^{n_1} X_i^2}{\sum_{j=n_1+1}^{n_1+n_2} X_j^2} \frac{n_2}{n_1}$$

then $F \sim \mathcal{F}(n_1, n_2)$. (See Table 8.4 for quantiles.)

Definition 8.12 (Student-t distribution)**.** Z follows a Student-t (or t) distribution with ν degrees of freedom (d.f.) if:

$$Z = X_0 \Big/ \sqrt{\frac{\sum_{i=1}^{\nu} X_i^2}{\nu}}, \quad X_i \sim i.i.d. \mathcal{N}(0,1).$$

We have $\mathbb{E}(Z) = 0$, and $\mathbb{V}ar(Z) = \frac{\nu}{\nu-2}$ if $\nu > 2$. (See Table 8.2 for quantiles.)

Definition 8.13 (Chi-square distribution)**.** Z follows a χ^2 distribution with ν d.f. if $Z = \sum_{i=1}^{\nu} X_i^2$ where $X_i \sim i.i.d. \mathcal{N}(0,1)$. We have $\mathbb{E}(Z) = \nu$. (See Table 8.3 for quantiles.)

Definition 8.14 (Cauchy distribution)**.** The probability distribution function of the Cauchy distribution defined by a location parameter μ and a scale parameter γ is:

$$f(x) = \frac{1}{\pi\gamma \left(1 + \left[\frac{x-\mu}{\gamma}\right]^2\right)}.$$

The mean and variance of this distribution are undefined.

Proposition 8.10 (Inner product of a multivariate Gaussian variable)**.** *Let X be a n-dimensional multivariate Gaussian variable: $X \sim \mathcal{N}(0, \Sigma)$. We have:*

$$X'\Sigma^{-1}X \sim \chi^2(n).$$

Proof. Because Σ is a symmetrical definite positive matrix, it admits the spectral decomposition PDP' where P is an orthogonal matrix (i.e. $PP' = Id$) and D is a diagonal matrix with non-negative entries. Denoting by $\sqrt{D^{-1}}$ the diagonal matrix whose diagonal entries are the inverse of those of D, it is

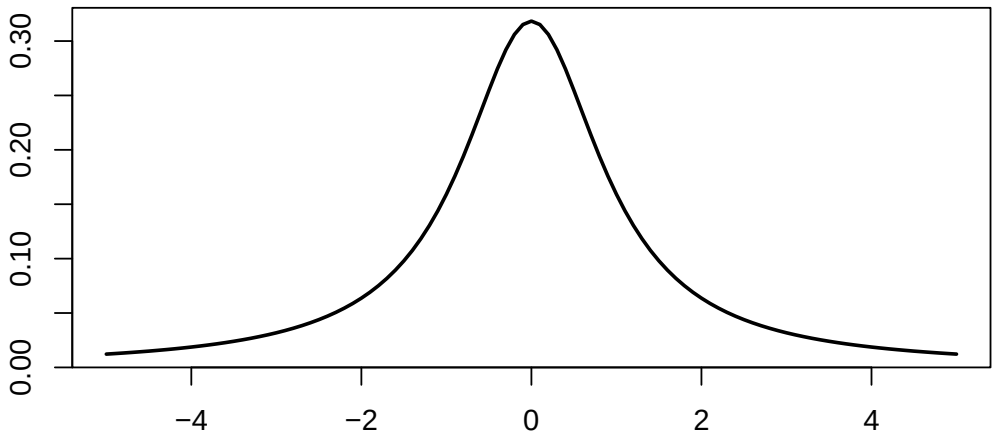


Figure 8.2: Pdf of the Cauchy distribution ($\mu = 0$, $\gamma = 1$).

easily checked that the covariance matrix of $Y := \sqrt{D^{-1}}P'X$ is Id. Therefore Y is a vector of uncorrelated Gaussian variables. The properties of Gaussian variables imply that the components of Y are then also independent. Hence $Y'Y = \sum_i Y_i^2 \sim \chi^2(n)$.

It remains to note that $Y'Y = X'PD^{-1}P'X = X'\mathbb{V}ar(X)^{-1}X$ to conclude. □

Definition 8.15 (Generalized Extreme Value (GEV) distribution)**.** The vector of disturbances $\varepsilon = [\varepsilon_{1,1}, \dots, \varepsilon_{1,K_1}, \dots, \varepsilon_{J,1}, \dots, \varepsilon_{J,K_J}]'$ follows the Generalized Extreme Value (GEV) distribution if its c.d.f. is:

$$F(\varepsilon, \rho) = \exp(-G(e^{-\varepsilon_{1,1}}, \dots, e^{-\varepsilon_{J,K_J}}; \rho))$$

with

$$\begin{aligned} G(\mathbf{Y}; \rho) &\equiv G(Y_{1,1}, \dots, Y_{1,K_1}, \dots, Y_{J,1}, \dots, Y_{J,K_J}; \rho) \\ &= \sum_{j=1}^{J} \left(\sum_{k=1}^{K_j} Y_{jk}^{1/\rho_j} \right)^{\rho_j} \end{aligned}$$

8.3.3 Stochastic convergences

Proposition 8.11 (Chebychev's inequality)**.** *If* $\mathbb{E}(|X|^r)$ *is finite for some* $r > 0$ *then:*

$$\forall \varepsilon > 0, \quad \mathbb{P}(|X - c| > \varepsilon) \leq \frac{\mathbb{E}[|X - c|^r]}{\varepsilon^r}.$$

In particular, for $r = 2$*:*

$$\forall \varepsilon > 0, \quad \mathbb{P}(|X - c| > \varepsilon) \leq \frac{\mathbb{E}[(X - c)^2]}{\varepsilon^2}.$$

Proof. Remark that $\varepsilon^r \mathbb{1}_{\{|X| \geq \varepsilon\}} \leq |X|^r$ and take the expectation of both sides. □

Definition 8.16 (Convergence in probability)**.** The random variable sequence x_n converges in probability to a constant c if $\forall \varepsilon$, $\lim_{n \to \infty} \mathbb{P}(|x_n - c| > \varepsilon) = 0$.

It is denoted as: $\operatorname{plim} x_n = c$.

Definition 8.17 (Convergence in the Lr norm)**.** x_n converges in the r-th mean (or in the L^r-norm) towards x, if $\mathbb{E}(|x_n|^r)$ and $\mathbb{E}(|x|^r)$ exist and if

$$\lim_{n \to \infty} \mathbb{E}(|x_n - x|^r) = 0.$$

It is denoted as: $x_n \xrightarrow{L^r} c$.

For $r = 2$, this convergence is called **mean square convergence**.

Definition 8.18 (Almost sure convergence)**.** The random variable sequence x_n converges almost surely to c if $\mathbb{P}(\lim_{n \to \infty} x_n = c) = 1$.

It is denoted as: $x_n \xrightarrow{a.s.} c$.

Definition 8.19 (Convergence in distribution)**.** x_n is said to converge in distribution (or in law) to x if

$$\lim_{n \to \infty} F_{x_n}(s) = F_x(s)$$

for all s at which F_X –the cumulative distribution of X– is continuous.

It is denoted as: $x_n \xrightarrow{d} x$.

Proposition 8.12 (Rules for limiting distributions (Slutsky)). *We have:*

i. ***Slutsky's theorem:*** *If* $x_n \xrightarrow{d} x$ *and* $y_n \xrightarrow{p} c$ *then*

$$\begin{aligned} x_n y_n &\xrightarrow{d} xc \\ x_n + y_n &\xrightarrow{d} x + c \\ x_n / y_n &\xrightarrow{d} x/c \quad (\text{if } c \neq 0) \end{aligned}$$

ii. ***Continuous mapping theorem:*** *If* $x_n \xrightarrow{d} x$ *and* g *is a continuous function then* $g(x_n) \xrightarrow{d} g(x)$.

Proposition 8.13 (Implications of stochastic convergences). *We have:*

$$\boxed{\xrightarrow{L^s}} \underset{1\leq r\leq s}{\Rightarrow} \boxed{\xrightarrow{L^r}} \quad \Downarrow \quad \boxed{\xrightarrow{a.s.}} \Rightarrow \boxed{\xrightarrow{p}} \Rightarrow \boxed{\xrightarrow{d}}.$$

Proof. (of the fact that $\left(\xrightarrow{p}\right) \Rightarrow \left(\xrightarrow{d}\right)$). Assume that $X_n \xrightarrow{p} X$. Denoting by F and F_n the c.d.f. of X and X_n, respectively:

$$\begin{aligned} F_n(x) &= \mathbb{P}(X_n \leq x, X \leq x+\varepsilon) + \mathbb{P}(X_n \leq x, X > x+\varepsilon) \\ &\leq F(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon). \end{aligned}$$

Besides,

$$\begin{aligned} F(x-\varepsilon) &= \mathbb{P}(X \leq x-\varepsilon, X_n \leq x) + \mathbb{P}(X \leq x-\varepsilon, X_n > x) \\ &\leq F_n(x) + \mathbb{P}(|X_n - X| > \varepsilon), \end{aligned}$$

which implies:

$$F(x-\varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \leq F_n(x). \tag{8.3}$$

Eqs. (8.3) and (8.3) imply:

$$F(x-\varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \leq F_n(x) \leq F(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).$$

Taking limits as $n \to \infty$ yields

$$F(x-\varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x+\varepsilon).$$

The result is then obtained by taking limits as $\varepsilon \to 0$ (if F is continuous at x). □

Proposition 8.14 (Convergence in distribution to a constant)**.** *If X_n converges in distribution to a constant c, then X_n converges in probability to c.*

Proof. If $\varepsilon > 0$, we have $\mathbb{P}(X_n < c-\varepsilon) \underset{n\to\infty}{\to} 0$ i.e. $\mathbb{P}(X_n \geq c-\varepsilon) \underset{n\to\infty}{\to} 1$ and $\mathbb{P}(X_n < c+\varepsilon) \underset{n\to\infty}{\to} 1$. Therefore $\mathbb{P}(c-\varepsilon \leq X_n < c+\varepsilon) \underset{n\to\infty}{\to} 1$, which gives the result. □

Example 8.1 (Convergence in probability but not L^r)**.** Let $\{x_n\}_{n\in\mathbb{N}}$ be a series of random variables defined by:

$$x_n = nu_n,$$

where u_n are independent random variables s.t. $u_n \sim \mathcal{B}(1/n)$.

We have $x_n \overset{p}{\to} 0$ but $x_n \overset{L^r}{\nrightarrow} 0$ because $\mathbb{E}(|X_n - 0|) = \mathbb{E}(X_n) = 1$.

Theorem 8.2 (Cauchy criterion (non-stochastic case))**.** *We have that $\sum_{i=0}^{T} a_i$ converges ($T \to \infty$) iff, for any $\eta > 0$, there exists an integer N such that, for all $M \geq N$,*

$$\left| \sum_{i=N+1}^{M} a_i \right| < \eta.$$

Theorem 8.3 (Cauchy criterion (stochastic case))**.** *We have that $\sum_{i=0}^{T} \theta_i \varepsilon_{t-i}$ converges in mean square ($T \to \infty$) to a random variable iff, for any $\eta > 0$, there exists an integer N such that, for all $M \geq N$,*

$$\mathbb{E}\left[\left(\sum_{i=N+1}^{M} \theta_i \varepsilon_{t-i} \right)^2 \right] < \eta.$$

8.3.4 Central limit theorem

Theorem 8.4 (Law of large numbers)**.** *The sample mean is a consistent estimator of the population mean.*

Proof. Let's denote by ϕ_{X_i} the characteristic function of a r.v. X_i. If the mean of X_i is μ then the Talyor expansion of the characteristic function is:

$$\phi_{X_i}(u) = \mathbb{E}(\exp(iuX)) = 1 + iu\mu + o(u).$$

The properties of the characteristic function (see Def. 8.10) imply that:

$$\phi_{\frac{1}{n}(X_1+\cdots+X_n)}(u) = \prod_{i=1}^{n}\left(1 + i\frac{u}{n}\mu + o\left(\frac{u}{n}\right)\right) \to e^{iu\mu}.$$

The facts that (a) $e^{iu\mu}$ is the characteristic function of the constant μ and (b) that a characteristic function uniquely characterises a distribution imply that the sample mean converges in distribution to the constant μ, which further implies that it converges in probability to μ. □

Theorem 8.5 (Lindberg-Levy Central limit theorem, CLT)**.** *If x_n is an i.i.d. sequence of random variables with mean μ and variance σ^2 ($\in]0, +\infty[$), then:*

$$\boxed{\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad \textit{where} \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.}$$

Proof. Let us introduce the r.v. $Y_n := \sqrt{n}(\bar{X}_n - \mu)$. We have $\phi_{Y_n}(u) = \left[\mathbb{E}\left(\exp(i\frac{1}{\sqrt{n}}u(X_1 - \mu))\right)\right]^n$. We have:

$$\begin{aligned}
& \left[\mathbb{E}\left(\exp\left(i\frac{1}{\sqrt{n}}u(X_1 - \mu)\right)\right)\right]^n \\
= & \left[\mathbb{E}\left(1 + i\frac{1}{\sqrt{n}}u(X_1 - \mu) - \frac{1}{2n}u^2(X_1 - \mu)^2 + o(u^2)\right)\right]^n \\
= & \left(1 - \frac{1}{2n}u^2\sigma^2 + o(u^2)\right)^n.
\end{aligned}$$

Therefore $\phi_{Y_n}(u) \underset{n\to\infty}{\to} \exp\left(-\frac{1}{2}u^2\sigma^2\right)$, which is the characteristic function of $\mathcal{N}(0, \sigma^2)$. □

8.4 Some properties of Gaussian variables

Proposition 8.15. *If* $\mathbf{A}$ *is idempotent and if* $\mathbf{x}$ *is Gaussian,* $\mathbf{Lx}$ *and* $\mathbf{x}'\mathbf{Ax}$ *are independent if* $\mathbf{LA} = \mathbf{0}$.

Proof. If $\mathbf{LA} = \mathbf{0}$, then the two Gaussian vectors $\mathbf{Lx}$ and $\mathbf{Ax}$ are independent. This implies the independence of any function of $\mathbf{Lx}$ and any function of $\mathbf{Ax}$. The results then follows from the observation that $\mathbf{x}'\mathbf{Ax} = (\mathbf{Ax})'(\mathbf{Ax})$, which is a function of $\mathbf{Ax}$. □

Proposition 8.16 (Bayesian update in a vector of Gaussian variables)**.** *If*

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}\right),$$

then

$$Y_2|Y_1 \sim \mathcal{N}\left(\Omega_{21}\Omega_{11}^{-1}Y_1, \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}\right).$$
$$Y_1|Y_2 \sim \mathcal{N}\left(\Omega_{12}\Omega_{22}^{-1}Y_2, \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right).$$

Proposition 8.17 (Truncated distributions)**.** *If* X *is a random variable distributed according to some p.d.f.* f*, with c.d.f.* F*, with infinite support. Then the p.d.f. of* $X|a \leq X < b$ *is*

$$g(x) = \frac{f(x)}{F(b) - F(a)}\mathbb{I}_{\{a \leq x < b\}},$$

for any $a < b$.

In partiucular, for a Gaussian variable $X \sim \mathcal{N}(\mu, \sigma^2)$*, we have*

$$f(X = x|a \leq X < b) = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{Z}.$$

with $Z = \Phi(\beta) - \Phi(\alpha)$*, where* $\alpha = \dfrac{a-\mu}{\sigma}$ *and* $\beta = \dfrac{b-\mu}{\sigma}$.

Moreover:

$$\mathbb{E}(X|a \leq X < b) = \mu - \frac{\phi(\beta) - \phi(\alpha)}{Z}\sigma. \tag{8.4}$$

We also have:

$$\begin{aligned} & \mathbb{V}ar(X|a \leq X < b) \\ = \; & \sigma^2 \left[1 - \frac{\beta\phi(\beta) - \alpha\phi(\alpha)}{Z} - \left(\frac{\phi(\beta) - \phi(\alpha)}{Z} \right)^2 \right] \end{aligned} \tag{8.5}$$

In particular, for $b \to \infty$, we get:

$$\mathbb{V}ar(X|a < X) = \sigma^2 \left[1 + \alpha\lambda(-\alpha) - \lambda(-\alpha)^2 \right], \tag{8.6}$$

*with $\lambda(x) = \dfrac{\phi(x)}{\Phi(x)}$ is called the **inverse Mills ratio**.*

Consider the case where $a \to -\infty$ (i.e. the conditioning set is $X < b$) and $\mu = 0$, $\sigma = 1$. Then Eq. (8.4) gives $\mathbb{E}(X|X < b) = -\lambda(b) = -\dfrac{\phi(b)}{\Phi(b)}$, where λ is the function computing the inverse Mills ratio.

Proposition 8.18 (p.d.f. of a multivariate Gaussian variable)**.** *If $Y \sim \mathcal{N}(\mu, \Omega)$ and if Y is a n-dimensional vector, then the density function of Y is:*

$$\frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \exp\left[-\frac{1}{2} (Y - \mu)' \, \Omega^{-1} (Y - \mu) \right].$$

8.5 Proofs

Proof of Proposition ??

Proof. Assumptions (i) and (ii) (in the set of Assumptions **??**) imply that θ_{MLE} exists ($= \text{argmax}_\theta (1/n) \log \mathcal{L}(\theta; \mathbf{y})$).

$(1/n) \log \mathcal{L}(\theta; \mathbf{y})$ can be interpreted as the sample mean of the r.v. $\log f(Y_i; \theta)$ that are i.i.d. Therefore $(1/n) \log \mathcal{L}(\theta; \mathbf{y})$ converges to $\mathbb{E}_{\theta_0}(\log f(Y; \theta))$ – which exists (Assumption iv).

Because the latter convergence is uniform (Assumption v), the solution θ_{MLE} almost surely converges to the solution to the limit problem:

$$\text{argmax}_\theta \mathbb{E}_{\theta_0}(\log f(Y; \theta)) = \text{argmax}_\theta \int_y \log f(y; \theta) f(y; \theta_0) dy.$$

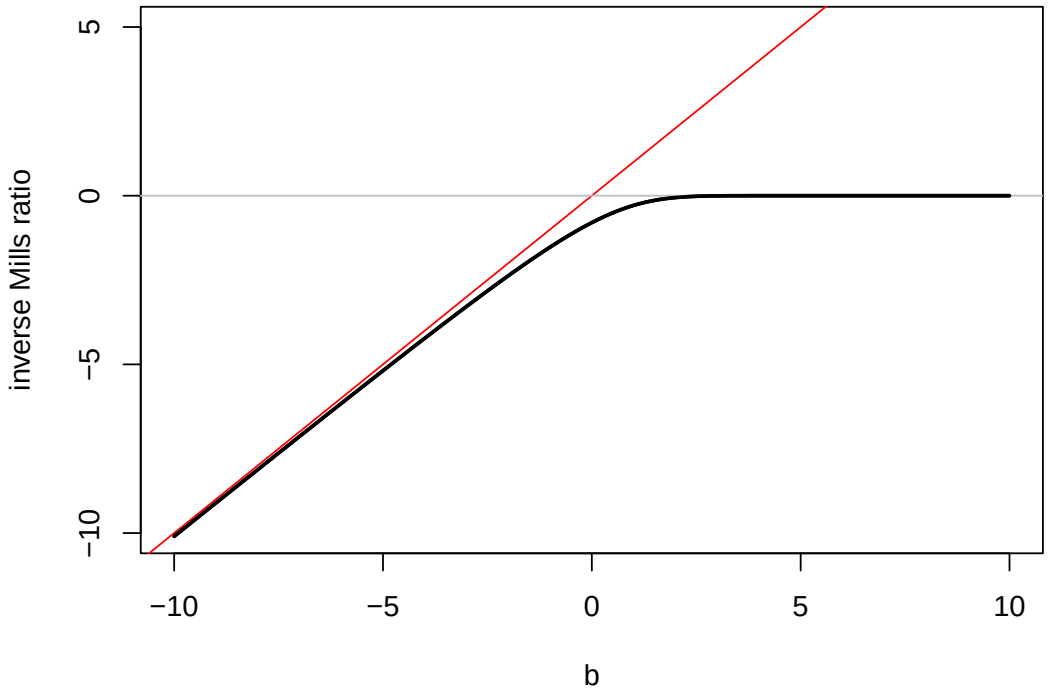


Figure 8.3: $\mathbb{E}(X|X < b)$ as a function of b when $X \sim \mathcal{N}(0, 1)$ (in black).

Properties of the Kullback information measure (see Prop. 8.9), together with the identifiability assumption (ii) implies that the solution to the limit problem is unique and equal to θ_0.

Consider a r.v. sequence θ that converges to θ_0. The Taylor expansion of the score in a neighborood of θ_0 yields to:

$$\frac{\partial \log \mathcal{L}(\theta;\mathbf{y})}{\partial \theta} = \frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} + \frac{\partial^2 \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta \partial \theta'}(\theta - \theta_0) + o_p(\theta - \theta_0)$$

θ_{MLE} converges to θ_0 and satisfies the likelihood equation $\frac{\partial \log \mathcal{L}(\theta;\mathbf{y})}{\partial \theta} = \mathbf{0}$. Therefore:

$$\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} \approx -\frac{\partial^2 \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta \partial \theta'}(\theta_{MLE} - \theta_0),$$

or equivalently:

$$\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} \approx \left(-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(y_i;\theta_0)}{\partial \theta \partial \theta'} \right) \sqrt{n}(\theta_{MLE} - \theta_0),$$

By the law of large numbers, we have: $\left(-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(y_i;\theta_0)}{\partial \theta \partial \theta'} \right) \to \frac{1}{n}\mathbf{I}(\theta_0) = \mathcal{I}_Y(\theta_0)$.

Besides, we have:

$$\begin{aligned} \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} &= \sqrt{n}\left(\frac{1}{n}\sum_i \frac{\partial \log f(y_i;\theta_0)}{\partial \theta} \right) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_i \left\{ \frac{\partial \log f(y_i;\theta_0)}{\partial \theta} - \mathbb{E}_{\theta_0}\frac{\partial \log f(Y_i;\theta_0)}{\partial \theta} \right\} \right) \end{aligned}$$

which converges to $\mathcal{N}(0, \mathcal{I}_Y(\theta_0))$ by the CLT.

Collecting the preceding results leads to (b). The fact that θ_{MLE} achieves the FDCR bound proves (c). □

Proof of Proposition ??

Proof. We have $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$ (Eq. **??**eq:normMLE)). A Taylor expansion around θ_0 yields to:

$$\sqrt{n}(h(\hat{\theta}_n) - h(\theta_0)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h(\theta_0)'}{\partial \theta}\right). \tag{8.7}$$

Under H_0, $h(\theta_0) = 0$ therefore:

$$\sqrt{n}h(\hat{\theta}_n) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h(\theta_0)'}{\partial \theta}\right). \tag{8.8}$$

Hence

$$\sqrt{n}\left(\frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h(\theta_0)'}{\partial \theta}\right)^{-1/2} h(\hat{\theta}_n) \xrightarrow{d} \mathcal{N}(0, Id).$$

Taking the quadratic form, we obtain:

$$nh(\hat{\theta}_n)'\left(\frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h(\theta_0)'}{\partial \theta}\right)^{-1} h(\hat{\theta}_n) \xrightarrow{d} \chi^2(r).$$

The fact that the test has asymptotic level α directly stems from what precedes. **Consistency of the test**: Consider $\theta_0 \in \Theta$. Because the MLE is consistent, $h(\hat{\theta}_n)$ converges to $h(\theta_0) \neq 0$. Eq. (8.7) is still valid. It implies that ξ_n^W converges to $+\infty$ and therefore that $\mathbb{P}_\theta(\xi_n^W \geq \chi^2_{1-\alpha}(r)) \to 1$. □

Proof of Proposition ??

Proof. Notations: "$\approx$" means "equal up to a term that converges to 0 in probability". We are under H_0. $\hat{\theta}^0$ is the constrained ML estimator; $\hat{\theta}$ denotes the unconstrained one.

We combine the two Taylor expansion: $h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'}(\hat{\theta}_n - \theta_0)$ and $h(\hat{\theta}_n^0) \approx \frac{\partial h(\theta_0)}{\partial \theta'}(\hat{\theta}_n^0 - \theta_0)$ and we use $h(\hat{\theta}_n^0) = 0$ (by definition) to get:

$$\sqrt{n}h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'} \sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0). \tag{8.9}$$

Besides, we have (using the definition of the information matrix):

$$\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta} \approx \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \tag{8.10}$$

and:

$$0 = \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n;\mathbf{y})}{\partial \theta} \approx \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{8.11}$$

Taking the difference and multiplying by $\mathcal{I}(\theta_0)^{-1}$:

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0) \approx \mathcal{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta}\mathcal{I}(\theta_0). \tag{8.12}$$

Eqs. (8.9) and (8.12) yield to:

$$\sqrt{n}h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta}. \tag{8.13}$$

Recall that $\hat{\theta}_n^0$ is the MLE of θ_0 under the constraint $h(\theta) = 0$. The vector of Lagrange multipliers $\hat{\lambda}_n$ associated to this program satisfies:

$$\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta} + \frac{\partial h'(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta}\hat{\lambda}_n = 0. \tag{8.14}$$

Substituting the latter equation in Eq. (8.13) gives:

$$\begin{aligned}\sqrt{n}h(\hat{\theta}_n) &\approx -\frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h'(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}} \\ &\approx -\frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h'(\theta_0;\mathbf{y})}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}},\end{aligned}$$

which yields:

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \approx -\left(\frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h'(\theta_0;\mathbf{y})}{\partial \theta}\right)^{-1}\sqrt{n}h(\hat{\theta}_n). \tag{8.15}$$

It follows, from Eq. (8.8), that:

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h'(\theta_0; \mathbf{y})}{\partial \theta}\right)^{-1}\right).$$

Taking the quadratic form of the last equation gives:

$$\frac{1}{n}\hat{\lambda}_n' \frac{\partial h(\hat{\theta}_n^0)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n^0)^{-1} \frac{\partial h'(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \hat{\lambda}_n \xrightarrow{d} \chi^2(r).$$

Using Eq. (8.14), it appears that the left-hand side term of the last equation is ξ^{LM} as defined in Eq. (**??**). Consistency: see Remark 17.3 in Gouriéroux and Monfort (1995). □

Proof of Proposition ??

Proof. Let us first demonstrate the asymptotic equivalence of ξ^{LM} and ξ^{LR}.

The second-order Taylor expansions of $\log \mathcal{L}(\hat{\theta}_n^0, \mathbf{y})$ and $\log \mathcal{L}(\hat{\theta}_n, \mathbf{y})$ are:

$$\begin{aligned}
\log \mathcal{L}(\hat{\theta}_n, \mathbf{y}) &\approx \log \mathcal{L}(\theta_0, \mathbf{y}) + \frac{\partial \log \mathcal{L}(\theta_0, \mathbf{y})}{\partial \theta'}(\hat{\theta}_n - \theta_0) \\
&\quad -\frac{n}{2}(\hat{\theta}_n - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \\
\log \mathcal{L}(\hat{\theta}_n^0, \mathbf{y}) &\approx \log \mathcal{L}(\theta_0, \mathbf{y}) + \frac{\partial \log \mathcal{L}(\theta_0, \mathbf{y})}{\partial \theta'}(\hat{\theta}_n^0 - \theta_0) \\
&\quad -\frac{n}{2}(\hat{\theta}_n^0 - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0).
\end{aligned}$$

Taking the difference, we obtain:

$$\begin{aligned}
\xi_n^{LR} &\approx 2\frac{\partial \log \mathcal{L}(\theta_0, \mathbf{y})}{\partial \theta'}(\hat{\theta}_n - \hat{\theta}_n^0) + n(\hat{\theta}_n^0 - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0) \\
&\quad -n(\hat{\theta}_n - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0).
\end{aligned}$$

Using $\frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta} \approx \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0)$ (Eq. (8.11)), we have:

$$\begin{aligned}
\xi_n^{LR} &\approx 2n(\hat{\theta}_n - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0) + n(\hat{\theta}_n^0 - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0) \\
&\quad -n(\hat{\theta}_n - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0).
\end{aligned}$$

In the second of the three terms in the sum, we replace $(\hat{\theta}_n^0 - \theta_0)$ by $(\hat{\theta}_n^0 - \hat{\theta}_n + \hat{\theta}_n - \theta_0)$ and we develop the associated product. This leads to:

$$\xi_n^{LR} \approx n(\hat{\theta}_n^0 - \hat{\theta}_n)' \mathcal{I}(\theta_0)^{-1} (\hat{\theta}_n^0 - \hat{\theta}_n). \tag{8.16}$$

The difference between Eqs. (8.10) and (8.11) implies:

$$\frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \approx \mathcal{I}(\theta_0) \sqrt{n} (\hat{\theta}_n - \hat{\theta}_n^0),$$

which, associated to Eq. @**??**eq:lr10), gives:

$$\xi_n^{LR} \approx \frac{1}{n} \frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \approx \xi_n^{LM}.$$

Hence ξ_n^{LR} has the same asymptotic distribution as ξ_n^{LM}.

Let's show that the LR test is consistent. For this, note that:

$$\begin{aligned} \frac{\log \mathcal{L}(\hat{\theta}, \mathbf{y}) - \log \mathcal{L}(\hat{\theta}^0, \mathbf{y})}{n} &= \frac{1}{n} \sum_{i=1}^{n} [\log f(y_i; \hat{\theta}_n) - \log f(y_i; \hat{\theta}_n^0)] \\ &\to \mathbb{E}_0[\log f(Y; \theta_0) - \log f(Y; \theta_\infty)], \end{aligned}$$

where θ_∞, the pseudo true value, is such that $h(\theta_\infty) \neq 0$ (by definition of H_1). From the Kullback inequality and the asymptotic identifiability of θ_0, it follows that $\mathbb{E}_0[\log f(Y; \theta_0) - \log f(Y; \theta_\infty)] > 0$. Therefore $\xi_n^{LR} \to +\infty$ under H_1.

Let us now demonstrate the equivalence of $\xi^{LM} and \xi^W$.

We have (using Eq. **??**eq:multiplier)):

$$\xi_n^{LM} = \frac{1}{n} \hat{\lambda}_n' \frac{\partial h(\hat{\theta}_n^0)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n^0)^{-1} \frac{\partial h'(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \hat{\lambda}_n.$$

Since, under H_0, $\hat{\theta}_n^0 \approx \hat{\theta}_n \approx \theta_0$, Eq. (8.15) therefore implies that:

$$\xi^{LM} \approx n h(\hat{\theta}_n)' \left(\frac{\partial h(\hat{\theta}_n)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n)^{-1} \frac{\partial h'(\hat{\theta}_n; \mathbf{y})}{\partial \theta} \right)^{-1} h(\hat{\theta}_n) = \xi^W,$$

which gives the result. □

Proof of Eq. (1.3)

Proof. We have:

$$\begin{aligned}
& T\mathbb{E}\left[(\bar{y}_T-\mu)^2\right] \\
= \; & T\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T}(y_t-\mu)\right)^2\right] = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(y_t-\mu)^2+2\sum_{s<t\leq T}(y_t-\mu)(y_s-\mu)\right] \\
= \; & \gamma_0+\frac{2}{T}\left(\sum_{t=2}^{T}\mathbb{E}\left[(y_t-\mu)(y_{t-1}-\mu)\right]\right)+\frac{2}{T}\left(\sum_{t=3}^{T}\mathbb{E}\left[(y_t-\mu)(y_{t-2}-\mu)\right]\right)+... \\
& +\frac{2}{T}\left(\sum_{t=T-1}^{T}\mathbb{E}\left[(y_t-\mu)(y_{t-(T-2)}-\mu)\right]\right)+\frac{2}{T}\mathbb{E}\left[(y_t-\mu)(y_{t-(T-1)}-\mu)\right] \\
= \; & \gamma_0+2\frac{T-1}{T}\gamma_1+\cdots+2\frac{1}{T}\gamma_{T-1}.
\end{aligned}$$

Therefore:

$$\begin{aligned}
& T\mathbb{E}\left[(\bar{y}_T-\mu)^2\right]-\sum_{j=-\infty}^{+\infty}\gamma_j \\
= \; & -2\frac{1}{T}\gamma_1-2\frac{2}{T}\gamma_2-\cdots-2\frac{T-1}{T}\gamma_{T-1}-2\gamma_T-2\gamma_{T+1}+...
\end{aligned}$$

And then:

$$\begin{aligned}
& \left|T\mathbb{E}\left[(\bar{y}_T-\mu)^2\right]-\sum_{j=-\infty}^{+\infty}\gamma_j\right| \\
\leq \; & 2\frac{1}{T}|\gamma_1|+2\frac{2}{T}|\gamma_2|+\cdots+2\frac{T-1}{T}|\gamma_{T-1}|+2|\gamma_T|+2|\gamma_{T+1}|+...
\end{aligned}$$

For any $q\leq T$, we have:

$$\begin{aligned}
\left|T\mathbb{E}\left[(\bar{y}_T-\mu)^2\right]-\sum_{j=-\infty}^{+\infty}\gamma_j\right| \; \leq \; & 2\frac{1}{T}|\gamma_1|+2\frac{2}{T}|\gamma_2|+\cdots+2\frac{q-1}{T}|\gamma_{q-1}|+2\frac{q}{T}|\gamma_q|+ \\
& 2\frac{q+1}{T}|\gamma_{q+1}|+\cdots+2\frac{T-1}{T}|\gamma_{T-1}|+2|\gamma_T|+2|\gamma_{T+1}|+... \\
\leq \; & \frac{2}{T}\left(|\gamma_1|+2|\gamma_2|+\cdots+(q-1)|\gamma_{q-1}|+q|\gamma_q|\right)+ \\
& 2|\gamma_{q+1}|+\cdots+2|\gamma_{T-1}|+2|\gamma_T|+2|\gamma_{T+1}|+...
\end{aligned}$$

Consider $\varepsilon > 0$. The fact that the autocovariances are absolutely summable implies that there exists q_0 such that (Cauchy criterion, Theorem 8.2):

$$2|\gamma_{q_0+1}| + 2|\gamma_{q_0+2}| + 2|\gamma_{q_0+3}| + \cdots < \varepsilon/2.$$

Then, if $T > q_0$, it comes that:

$$\left| T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] - \sum_{j=-\infty}^{+\infty} \gamma_j \right| \leq \frac{2}{T}\left(|\gamma_1| + 2|\gamma_2| + \cdots + (q_0-1)|\gamma_{q_0-1}| + q_0|\gamma_{q_0}|\right) + \varepsilon/2.$$

If $T \geq 2\left(|\gamma_1| + 2|\gamma_2| + \cdots + (q_0-1)|\gamma_{q_0-1}| + q_0|\gamma_{q_0}|\right)/(\varepsilon/2)$ $(= f(q_0)$, say) then

$$\frac{2}{T}\left(|\gamma_1| + 2|\gamma_2| + \cdots + (q_0-1)|\gamma_{q_0-1}| + q_0|\gamma_{q_0}|\right) \leq \varepsilon/2.$$

Then, if $T > f(q_0)$ and $T > q_0$, i.e. if $T > \max(f(q_0), q_0)$, we have:

$$\left| T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] - \sum_{j=-\infty}^{+\infty} \gamma_j \right| \leq \varepsilon.$$

□

Proof of Proposition 4.1

Proof. We have:

$$\begin{aligned}
\mathbb{E}([y_{t+1} - y^*_{t+1}]^2) &= \mathbb{E}\left([\{y_{t+1} - \mathbb{E}(y_{t+1}|x_t)\} + \{\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}\}]^2\right) \\
&= \mathbb{E}\left([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]^2\right) + \mathbb{E}\left([\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]^2\right) \\
&\quad +2\mathbb{E}\left([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)][\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]\right). \quad (8.17)
\end{aligned}$$

Let us focus on the last term. We have:

$$\begin{aligned}
&\mathbb{E}\left([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)][\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]\right) \\
&= \mathbb{E}(\mathbb{E}([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]\underbrace{[\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]}_{\text{function of } x_t}|x_t)) \\
&= \mathbb{E}([\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]\mathbb{E}([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]|x_t)) \\
&= \mathbb{E}([\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]\underbrace{[\mathbb{E}(y_{t+1}|x_t) - \mathbb{E}(y_{t+1}|x_t)]}_{=0}) = 0.
\end{aligned}$$

Therefore, Eq. (8.17) becomes:

$$\begin{aligned}&\mathbb{E}([y_{t+1}-y^*_{t+1}]^2)\\ =\quad &\underbrace{\mathbb{E}\left([y_{t+1}-\mathbb{E}(y_{t+1}|x_t)]^2\right)}_{\geq 0 \text{ and does not depend on } y^*_{t+1}} + \underbrace{\mathbb{E}\left([\mathbb{E}(y_{t+1}|x_t)-y^*_{t+1}]^2\right)}_{\geq 0 \text{ and depends on } y^*_{t+1}}.\end{aligned}$$

This implies that $\mathbb{E}([y_{t+1}-y^*_{t+1}]^2)$ is always larger than $\mathbb{E}([y_{t+1}-\mathbb{E}(y_{t+1}|x_t)]^2)$, and is therefore minimized if the second term is equal to zero, that is if $\mathbb{E}(y_{t+1}|x_t)=y^*_{t+1}$. □

Proof of Proposition 3.1

Proof. Using Proposition **??** (in Appendix **??**), we obtain that, conditionally on x_1, the log-likelihood is given by

$$\begin{aligned}\log\mathcal{L}(Y_T;\theta) &= -(Tn/2)\log(2\pi)+(T/2)\log\left|\Omega^{-1}\right|\\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\left[(y_t-\Pi' x_t)'\,\Omega^{-1}\,(y_t-\Pi' x_t)\right].\end{aligned}$$

Let's rewrite the last term of the log-likelihood:

$$\begin{aligned}\sum_{t=1}^{T}\left[(y_t-\Pi' x_t)'\,\Omega^{-1}\,(y_t-\Pi' x_t)\right] &=\\ \sum_{t=1}^{T}\left[\left(y_t-\hat{\Pi}' x_t+\hat{\Pi}' x_t-\Pi' x_t\right)'\Omega^{-1}\left(y_t-\hat{\Pi}' x_t+\hat{\Pi}' x_t-\Pi' x_t\right)\right] &=\\ \sum_{t=1}^{T}\left[\left(\hat{\varepsilon}_t+(\hat{\Pi}-\Pi)' x_t\right)'\Omega^{-1}\left(\hat{\varepsilon}_t+(\hat{\Pi}-\Pi)' x_t\right)\right],\end{aligned}$$

where the j^{th} element of the $(n\times 1)$ vector $\hat{\varepsilon}_t$ is the sample residual, for observation t, from an OLS regression of $y_{j,t}$ on x_t. Expanding the previous equation, we get:

$$\begin{aligned}&\sum_{t=1}^{T}\left[(y_t-\Pi' x_t)'\,\Omega^{-1}\,(y_t-\Pi' x_t)\right]=\sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}\hat{\varepsilon}_t\\ &+2\sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}(\hat{\Pi}-\Pi)' x_t+\sum_{t=1}^{T}x_t'(\hat{\Pi}-\Pi)\Omega^{-1}(\hat{\Pi}-\Pi)' x_t.\end{aligned}$$

Let's apply the trace operator on the second term (that is a scalar):

$$\begin{aligned}\sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}(\hat{\Pi}-\Pi)'x_t &= Tr\left(\sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}(\hat{\Pi}-\Pi)'x_t\right)\\ = Tr\left(\sum_{t=1}^{T}\Omega^{-1}(\hat{\Pi}-\Pi)'x_t\hat{\varepsilon}_t'\right) &= Tr\left(\Omega^{-1}(\hat{\Pi}-\Pi)'\sum_{t=1}^{T}x_t\hat{\varepsilon}_t'\right).\end{aligned}$$

Given that, by construction (property of OLS estimates), the sample residuals are orthogonal to the explanatory variables, this term is zero. Introducing $\tilde{x}_t = (\hat{\Pi}-\Pi)'x_t$, we have

$$\sum_{t=1}^{T}\left[(y_t-\Pi'x_t)'\,\Omega^{-1}\,(y_t-\Pi'x_t)\right] = \sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}\hat{\varepsilon}_t + \sum_{t=1}^{T}\tilde{x}_t'\Omega^{-1}\tilde{x}_t.$$

Since Ω is a positive definite matrix, Ω^{-1} is as well. Consequently, the smallest value that the last term can take is obtained for $\tilde{x}_t = 0$, i.e. when $\Pi = \hat{\Pi}$.

The MLE of Ω is the matrix $\hat{\Omega}$ that maximizes $\Omega \xrightarrow{\ell} L(Y_T;\hat{\Pi},\Omega)$. We have:

$$\log\mathcal{L}(Y_T;\hat{\Pi},\Omega) = -(Tn/2)\log(2\pi) + (T/2)\log\left|\Omega^{-1}\right| - \frac{1}{2}\sum_{t=1}^{T}\left[\hat{\varepsilon}_t'\Omega^{-1}\hat{\varepsilon}_t\right].$$

Matrix $\hat{\Omega}$ is a symmetric positive definite. It is easily checked that the (unrestricted) matrix that maximizes the latter expression is symmetric positive definite matrix. Indeed:

$$\frac{\partial\log\mathcal{L}(Y_T;\hat{\Pi},\Omega)}{\partial\Omega} = \frac{T}{2}\Omega' - \frac{1}{2}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t' \Rightarrow \hat{\Omega}' = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t',$$

which leads to the result. □

Proof of Proposition 3.2

Proof. Let us drop the i subscript. Rearranging Eq. (3.12), we have:

$$\sqrt{T}(\mathbf{b}-\beta) = (X'X/T)^{-1}\sqrt{T}(X'\varepsilon/T).$$

Let us consider the autocovariances of $\mathbf{v}_t = x_t\varepsilon_t$, denoted by γ_j^v. Using the fact that x_t is a linear combination of past ε_ts and that ε_t is a white noise, we get that $\mathbb{E}(\varepsilon_t x_t) = 0$. Therefore

$$\gamma_j^v = \mathbb{E}(\varepsilon_t\varepsilon_{t-j}x_t x_{t-j}').$$

If $j > 0$, we have $\mathbb{E}(\varepsilon_t\varepsilon_{t-j}x_t x_{t-j}') = \mathbb{E}(\mathbb{E}[\varepsilon_t\varepsilon_{t-j}x_t x_{t-j}'|\varepsilon_{t-j}, x_t, x_{t-j}]) = \mathbb{E}(\varepsilon_{t-j}x_t x_{t-j}'\mathbb{E}[\varepsilon_t|\varepsilon_{t-j}, x_t, x_{t-j}]) = 0$. Note that we have $\mathbb{E}[\varepsilon_t|\varepsilon_{t-j}, x_t, x_{t-j}] = 0$ because $\{\varepsilon_t\}$ is an i.i.d. white noise sequence. If $j = 0$, we have:

$$\gamma_0^v = \mathbb{E}(\varepsilon_t^2 x_t x_t') = \mathbb{E}(\varepsilon_t^2)\mathbb{E}(x_t x_t') = \sigma^2\mathbf{Q}.$$

The convergence in distribution of $\sqrt{T}(X'\varepsilon/T) = \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} v_t$ results from the Central Limit Theorem for covariance-stationary processes, using the γ_j^v computed above. □

8.6 Additional codes

8.6.1 Simulating GEV distributions

The following lines of code have been used to generate Figure **??**.

```
n.sim <- 4000
par(mfrow=c(1,3),
    plt=c(.2,.95,.2,.85))
all.rhos <- c(.3,.6,.95)
for(j in 1:length(all.rhos)){
  theta <- 1/all.rhos[j]
  v1 <- runif(n.sim)
  v2 <- runif(n.sim)
  w <- rep(.000001,n.sim)
  # solve for f(w) = w*(1 - log(w)/theta) - v2 = 0
  for(i in 1:20){
    f.i <- w * (1 - log(w)/theta) - v2
    f.prime <- 1 - log(w)/theta - 1/theta
    w <- w - f.i/f.prime
  }
```

```
  u1 <- exp(v1^(1/theta) * log(w))
  u2 <- exp((1-v1)^(1/theta) * log(w))

  # Get eps1 and eps2 using the inverse of
  # the Gumbel distribution's cdf:
  eps1 <- -log(-log(u1))
  eps2 <- -log(-log(u2))
  cbind(cor(eps1,eps2),1-all.rhos[j]^2)
  plot(eps1,eps2,pch=19,col="#FF000044",
       main=paste("rho = ",toString(all.rhos[j]),sep=""),
       xlab=expression(epsilon[1]),
       ylab=expression(epsilon[2]),
       cex.lab=2,cex.main=1.5)
}
```

8.6.2 Computing the covariance matrix of IRF using the delta method

```
irf.function <- function(THETA){
  c <- THETA[1]
  phi <- THETA[2:(p+1)]
  if(q>0){
    theta <- c(1,THETA[(1+p+1):(1+p+q)])
  }else{
    theta <- 1
  }
  sigma <- THETA[1+p+q+1]
  r <- dim(Matrix.of.Exog)[2] - 1
  beta <- THETA[(1+p+q+1+1):(1+p+q+1+(r+1))]

  irf <- sim.arma(0,phi,beta,sigma=sd(Ramey$ED3_TC,na.rm=TRUE),T=60,
                  y.0=rep(0,length(x$phi)),nb.sim=1,make.IRF=1,
                  X=NaN,beta=NaN)
  return(irf)
}
```

```
IRF.0 <- 100*irf.function(x$THETA)
eps <- .00000001
d.IRF <- NULL
for(i in 1:length(x$THETA)){
  THETA.i <- x$THETA
  THETA.i[i] <- THETA.i[i] + eps
  IRF.i <- 100*irf.function(THETA.i)
  d.IRF <- cbind(d.IRF,
                 (IRF.i - IRF.0)/eps
                 )
}
mat.var.cov.IRF <- d.IRF %*% x$I %*% t(d.IRF)
```

8.7 Statistical Tables

Table 8.1: Quantiles of the $\mathcal{N}(0,1)$ distribution. If a and b are respectively the row and column number; then the corresponding cell gives $\mathbb{P}(0 < X \leq a+b)$, where $X \sim \mathcal{N}(0,1)$.

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.6179	0.7257	0.8159	0.8849	0.9332	0.9641	0.9821	0.9918	0.9965
0.1	0.5040	0.6217	0.7291	0.8186	0.8869	0.9345	0.9649	0.9826	0.9920	0.9966
0.2	0.5080	0.6255	0.7324	0.8212	0.8888	0.9357	0.9656	0.9830	0.9922	0.9967
0.3	0.5120	0.6293	0.7357	0.8238	0.8907	0.9370	0.9664	0.9834	0.9925	0.9968
0.4	0.5160	0.6331	0.7389	0.8264	0.8925	0.9382	0.9671	0.9838	0.9927	0.9969
0.5	0.5199	0.6368	0.7422	0.8289	0.8944	0.9394	0.9678	0.9842	0.9929	0.9970
0.6	0.5239	0.6406	0.7454	0.8315	0.8962	0.9406	0.9686	0.9846	0.9931	0.9971
0.7	0.5279	0.6443	0.7486	0.8340	0.8980	0.9418	0.9693	0.9850	0.9932	0.9972
0.8	0.5319	0.6480	0.7517	0.8365	0.8997	0.9429	0.9699	0.9854	0.9934	0.9973
0.9	0.5359	0.6517	0.7549	0.8389	0.9015	0.9441	0.9706	0.9857	0.9936	0.9974
1	0.5398	0.6554	0.7580	0.8413	0.9032	0.9452	0.9713	0.9861	0.9938	0.9974
1.1	0.5438	0.6591	0.7611	0.8438	0.9049	0.9463	0.9719	0.9864	0.9940	0.9975
1.2	0.5478	0.6628	0.7642	0.8461	0.9066	0.9474	0.9726	0.9868	0.9941	0.9976
1.3	0.5517	0.6664	0.7673	0.8485	0.9082	0.9484	0.9732	0.9871	0.9943	0.9977
1.4	0.5557	0.6700	0.7704	0.8508	0.9099	0.9495	0.9738	0.9875	0.9945	0.9977
1.5	0.5596	0.6736	0.7734	0.8531	0.9115	0.9505	0.9744	0.9878	0.9946	0.9978
1.6	0.5636	0.6772	0.7764	0.8554	0.9131	0.9515	0.9750	0.9881	0.9948	0.9979
1.7	0.5675	0.6808	0.7794	0.8577	0.9147	0.9525	0.9756	0.9884	0.9949	0.9979
1.8	0.5714	0.6844	0.7823	0.8599	0.9162	0.9535	0.9761	0.9887	0.9951	0.9980
1.9	0.5753	0.6879	0.7852	0.8621	0.9177	0.9545	0.9767	0.9890	0.9952	0.9981
2	0.5793	0.6915	0.7881	0.8643	0.9192	0.9554	0.9772	0.9893	0.9953	0.9981
2.1	0.5832	0.6950	0.7910	0.8665	0.9207	0.9564	0.9778	0.9896	0.9955	0.9982
2.2	0.5871	0.6985	0.7939	0.8686	0.9222	0.9573	0.9783	0.9898	0.9956	0.9982
2.3	0.5910	0.7019	0.7967	0.8708	0.9236	0.9582	0.9788	0.9901	0.9957	0.9983
2.4	0.5948	0.7054	0.7995	0.8729	0.9251	0.9591	0.9793	0.9904	0.9959	0.9984
2.5	0.5987	0.7088	0.8023	0.8749	0.9265	0.9599	0.9798	0.9906	0.9960	0.9984
2.6	0.6026	0.7123	0.8051	0.8770	0.9279	0.9608	0.9803	0.9909	0.9961	0.9985
2.7	0.6064	0.7157	0.8078	0.8790	0.9292	0.9616	0.9808	0.9911	0.9962	0.9985
2.8	0.6103	0.7190	0.8106	0.8810	0.9306	0.9625	0.9812	0.9913	0.9963	0.9986
2.9	0.6141	0.7224	0.8133	0.8830	0.9319	0.9633	0.9817	0.9916	0.9964	0.9986

Table 8.2: Quantiles of the Student-t distribution. The rows correspond to different degrees of freedom (ν, say); the columns correspond to different probabilities (z, say). The cell gives q that is s.t. $\mathbb{P}(-q < X < q) = z$, with $X \sim t(\nu)$.

	0.05	0.1	0.75	0.9	0.95	0.975	0.99	0.999
1	0.079	0.158	2.414	6.314	12.706	25.452	63.657	636.619
2	0.071	0.142	1.604	2.920	4.303	6.205	9.925	31.599
3	0.068	0.137	1.423	2.353	3.182	4.177	5.841	12.924
4	0.067	0.134	1.344	2.132	2.776	3.495	4.604	8.610
5	0.066	0.132	1.301	2.015	2.571	3.163	4.032	6.869
6	0.065	0.131	1.273	1.943	2.447	2.969	3.707	5.959
7	0.065	0.130	1.254	1.895	2.365	2.841	3.499	5.408
8	0.065	0.130	1.240	1.860	2.306	2.752	3.355	5.041
9	0.064	0.129	1.230	1.833	2.262	2.685	3.250	4.781
10	0.064	0.129	1.221	1.812	2.228	2.634	3.169	4.587
20	0.063	0.127	1.185	1.725	2.086	2.423	2.845	3.850
30	0.063	0.127	1.173	1.697	2.042	2.360	2.750	3.646
40	0.063	0.126	1.167	1.684	2.021	2.329	2.704	3.551
50	0.063	0.126	1.164	1.676	2.009	2.311	2.678	3.496
60	0.063	0.126	1.162	1.671	2.000	2.299	2.660	3.460
70	0.063	0.126	1.160	1.667	1.994	2.291	2.648	3.435
80	0.063	0.126	1.159	1.664	1.990	2.284	2.639	3.416
90	0.063	0.126	1.158	1.662	1.987	2.280	2.632	3.402
100	0.063	0.126	1.157	1.660	1.984	2.276	2.626	3.390
200	0.063	0.126	1.154	1.653	1.972	2.258	2.601	3.340
500	0.063	0.126	1.152	1.648	1.965	2.248	2.586	3.310

Table 8.3: Quantiles of the χ^2 distribution. The rows correspond to different degrees of freedom; the columns correspond to different probabilities.

	0.05	0.1	0.75	0.9	0.95	0.975	0.99	0.999
1	0.004	0.016	1.323	2.706	3.841	5.024	6.635	10.828
2	0.103	0.211	2.773	4.605	5.991	7.378	9.210	13.816
3	0.352	0.584	4.108	6.251	7.815	9.348	11.345	16.266
4	0.711	1.064	5.385	7.779	9.488	11.143	13.277	18.467
5	1.145	1.610	6.626	9.236	11.070	12.833	15.086	20.515
6	1.635	2.204	7.841	10.645	12.592	14.449	16.812	22.458
7	2.167	2.833	9.037	12.017	14.067	16.013	18.475	24.322
8	2.733	3.490	10.219	13.362	15.507	17.535	20.090	26.124
9	3.325	4.168	11.389	14.684	16.919	19.023	21.666	27.877
10	3.940	4.865	12.549	15.987	18.307	20.483	23.209	29.588
20	10.851	12.443	23.828	28.412	31.410	34.170	37.566	45.315
30	18.493	20.599	34.800	40.256	43.773	46.979	50.892	59.703
40	26.509	29.051	45.616	51.805	55.758	59.342	63.691	73.402
50	34.764	37.689	56.334	63.167	67.505	71.420	76.154	86.661
60	43.188	46.459	66.981	74.397	79.082	83.298	88.379	99.607
70	51.739	55.329	77.577	85.527	90.531	95.023	100.425	112.317
80	60.391	64.278	88.130	96.578	101.879	106.629	112.329	124.839
90	69.126	73.291	98.650	107.565	113.145	118.136	124.116	137.208
100	77.929	82.358	109.141	118.498	124.342	129.561	135.807	149.449
200	168.279	174.835	213.102	226.021	233.994	241.058	249.445	267.541
500	449.147	459.926	520.950	540.930	553.127	563.852	576.493	603.446

Table 8.4: Quantiles of the $\mathcal{F}$ distribution. The columns and rows correspond to different degrees of freedom (resp. n_1 and n_2). The different panels correspond to different probabilities (α) The corresponding cell gives z that is s.t. $\mathbb{P}(X \leq z) = \alpha$, with $X \sim \mathcal{F}(n_1, n_2)$.

	1	2	3	4	5	6	7	8	9
alpha = 0.9									
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965
50	2.809	2.412	2.197	2.061	1.966	1.895	1.840	1.796	1.760
100	2.756	2.356	2.139	2.002	1.906	1.834	1.778	1.732	1.695
500	2.716	2.313	2.095	1.956	1.859	1.786	1.729	1.683	1.644
alpha = 0.95									
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975
500	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899
alpha = 0.99									
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590
500	6.686	4.648	3.821	3.357	3.054	2.838	2.675	2.547	2.443

Bibliography

Abadir, K. M. (1993). Ols bias in a nonstationary autoregression. *Econometric Theory*, 9(1):81–93.

Anderson, T. (1971). *The Statistical Analysis of Time Series*. Wiley.

Bernanke, B. S. (1986). Alternative explanations of the money-income correlation. *Carnegie-Rochester Conference Series on Public Policy*, 25:49–99.

Blanchard, O. J. and Quah, D. (1989). The Dynamic Effects of Aggregate Demand and Supply Disturbances. *American Economic Review*, 79(4):655–673.

Box, G. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Christiano, L. J., Eichenbaum, M., and Evans, C. (1996). The effects of monetary policy shocks: Evidence from the flow of funds. *The Review of Economics and Statistics*, 78(1):16–34.

Christiano, L. J., Eichenbaum, M., and Vigfusson, R. (2007). *Assessing Structural VARs*, pages 1–106. MIT Press.

De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22(3):443–473. Twenty five years of forecasting.

Dedola, L. and Lippi, F. (2005). The monetary transmission mechanism: Evidence from the industries of five oecd countries. *European Economic Review*, 49(6):1543–1569.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431.

Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276.

Erceg, C. J., Guerrieri, L., and Gust, C. (2005). Can Long-Run Restrictions Identify Technology Shocks? *Journal of the European Economic Association*, 3(6):1237–1278.

Faust, J. and Leeper, E. M. (1997). When do long-run identifying restrictions give reliable results? *Journal of Business & Economic Statistics*, 15(3):345–353.

Fischer, S. (1977). Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy*, 85(1):191–205.

Galí, J. (1999). Technology, employment, and the business cycle: Do technology shocks explain aggregate fluctuations? *American Economic Review*, 89(1):249–271.

Galí, J. (1992). How well does the is-lm model fit postwar u.s. data? *The Quarterly Journal of Economics*, 107(2):709–738.

Gerlach, S. and Smets, F. (1995). The Monetary Transmission Mechanism: Evidence from the G-7 Countries. CEPR Discussion Papers 1219, C.E.P.R. Discussion Papers.

Gertler, M. and Karadi, P. (2015). Monetary Policy Surprises, Credit Costs, and Economic Activity. *American Economic Journal: Macroeconomics*, 7(1):44–76.

Gouriéroux, C. and Monfort, A. (1995). *Statistics and Econometric Models*, volume 1 of *Themes in Modern Econometrics*. Cambridge University Press.

Gouriéroux, C., Monfort, A., and Renne, J.-P. (2020). Identification and Estimation in Non-Fundamental Structural VARMA Models. *Review of Economic Studies*, 87(4):1915–1953.

Gouriéroux, C., Monfort, A., and Renne, J.-P. (2017). Statistical inference for independent component analysis: Application to structural var models. *Journal of Econometrics*, 196(1):111–126.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, 59(6):1551–1580.

Jordà, O., Schularick, M., and Taylor, A. M. (2017). Macrofinancial History and the New Business Cycle Facts. *NBER Macroeconomics Annual*, 31(1):213–263.

Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics*, 80(2):218–230.

Kim, J. H. (2022). *VAR.etp: VAR Modelling: Estimation, Testing, and Prediction*. R package version 1.0.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178.

Litterman, R. and Scheinkman, J. (1991). Common Factors Affecting Bond Returns. *Journal of Fixed Income*, (1):54–61.

Lütkepohl, H. (1990). Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *The Review of Economics and Statistics*, 72(1):116–25.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Phillips, P. C. B. (1987). Time series regression with a unit root. *Econometrica*, 55(2):277–301.

Phillips, P. C. B. and Ouliaris, S. (1990). Asymptotic properties of residual based tests for cointegration. *Econometrica*, 58(1):165–193.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.

Ramey, V. A. (2016). Macroeconomic Shocks and Their Propagation. NBER Working Papers 21978, National Bureau of Economic Research, Inc.

Ruibio-Ramírez, J. F., Waggoner, D. F., and Zha, T. (2010). Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies*, 77(2):665–696.

Schwert, G. W. (1989). Tests for Unit Roots: A Monte Carlo Investigation. *Journal of Business & Economic Statistics*, 7(2):147–159.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Sims, C. A. (1986). Are forecasting models usable for policy analysis? *Quarterly Review*, 10(Win):2–16.

Stock, J. and Watson, M. (2016). Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. In Taylor, J. B. and Uhlig, H., editors, *Handbook of Macroeconomics*, volume 2 of *Handbook of Macroeconomics*, chapter 0, pages 415–525. Elsevier.

Stock, J. and Watson, M. W. (2003). *Introduction to Econometrics*. Prentice Hall, New York.