A Graphical Modelling Approach to Time Series

Marco Reale

Submitted for the degree of Doctor of Philosophy
at Lancaster University,
September 1998.

To Patrizia

Abstract

In this thesis we apply graphical modelling to the identification of time series models, both for the univariate and the multivariate case.

Graphical modelling reveals conditional dependences among variables leading to more simple models of their structure. Its application to time series models requires special considerations. Firstly, the models we require must explain the causal structure and this must be derived from the conditional dependence structure. Secondly, when graphical modelling is applied to a single sample of a univariate or multivariate time series it is necessary to verify to what extent it is possible to apply statistical tests used in the case of independent replicated multivariate samples.

In this thesis we proceed to analyse the potential of graphical modelling in the time series context on the basis of these considerations.

In the univariate case the graphical modelling approach compares well with other more established methods. In particular we use a lemma of Ljung and Box to determine the conditional dependence structure for a univariate autoregressive model and compare this approach to the use of the partial autocorrelation function. In the context of multivariate AR models the strength of graphical modelling is that it directs attention to a relatively small number of structural AR models. This is achieved by the application of the inverse variance lemma and the procedure of *demoralisation* which reduce the very large number of models which would have to be considered.

In the final part of this thesis we apply graphical modelling to multivariate ARMA models. A procedure of Durbin for ARMA models estimation suggests an initial approach in the context. The efficiency of this is improved using a filtering operation encountered in maximum likelihood estimation of these models.

Throughout the thesis we apply the results to simulated and real data sets. The real

examples concern two important monetary economics investigations. The first is the *lending channel* monetary transmission mechanism in the Italian economy where we obtained statistical evidence of a direct but weak influence of repurchase agreement interest rate on the banking loan interest rate. The second is the term structure of the U.S. dollar interest rate where we establish that the two year rate is pivotal in determining the other period interest rates, from 6 months to ten years..

Acknowledgements

This is the page of my thesis which I am going to write with more pleasure as it makes me remember all the nice moments I had, really a lot, since I came to Lancaster.

This is the first opportunity I have to express formally my gratitude to Granville Tunnicliffe Wilson, my supervisor, for his patient guidance and friendship. I enjoyed our statistical conversations, our non statistical conversations and our walks near Yealand.

Then I would like to thank all the people in my department, Mathematics and Statistics, for their support and their friendly and informal attitude which made my studying at Lancaster easy and enjoyable. In particular I wish to thank Stuart Coles and Joe Whittaker for their useful comments.

For financial support I thank the *Foreign and Commonwealth Office* and the *British Council* for the award of the *British Chevening Scholarship* and also my department.

It is impossible to mention all the friends who shared these years with me. To all of them my deepest gratitude.

I owe my parents and relatives a great deal for their support. Among them my cousin, Patrizia, who is not with us anymore. This thesis is dedicated to her.

Finally I thank my wife, Brunella, for her patience, encouragement and love and her willingness to come with me to the other side of the world.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

My own interest in the subject of this thesis, time series, is in its application to the development of models of the economy. A very straightforward way to verify empirically an economic theory is to translate it into a mathematical language which would lead to one or more equations and estimating them. Different tests applied to these equations would possibly give an answer to whether the theory is reliable or not. This is, in a nutshell, econometrics which, in the words of the founders of the Econometric Society (Irving Fisher, Ragnar Frisch and Jan Tinbergen) aims, to the advancement of economic theory in its relation to statistics and mathematics. The point is that the formulation of the theory, i.e. the *structural equations* come before data analysis, which contributes to the estimation and testing of the equations, but not primarily to formulating their structure.

The development of computers from the late forties made it possible to estimate systems which offered a detailed description of the economies with a huge amount of equations. By solving these models for some of the variables we would reduce the number of equations going, in this way, from the structural to a *reduced* form of the model. Reduced forms are more parsimonious in parameters but loose in terms of descriptive power, i.e the meaning is not so apparent in the reduced form. For example suppose a model relates unemployment (u) to production (y). A higher level of production would reduce the unemployment which in turn, by increasing the aggregate demand, will positively affect the production; this is known as the *Keynesian multiplier*. On the other hand a higher level of unemployment would reduce the aggregate demand and then the production.

This mechanism is summarised by the system of two structural equations 1.1.1.

$$\begin{cases} u_t &= \alpha_1 + \alpha_2 y_{t+k-m} \\ y_t &= \beta u_{t-k}. \end{cases} \tag{1.1.1}$$

where the parameters can be either positive or negative.

We can obtain a description more parsimonious in parameters by solving 1.1.1 for u, obtaining in this way the reduced form

$$u_t = \gamma_1 + \gamma_2 u_{t-m} \tag{1.1.2}$$

where $\gamma_1 = \alpha_1$ and $\gamma_2 = \alpha_2\beta$. In general the reduced form models involve fewer variables and their lagged values.

In the sixties, time series models began to be developed, for example those of the ARIMA class. These had an appearance similar to the reduced form models, but their specification was carried out purely by statistical criteria without reference to preconceived economic theory.

Until the early seventies they were therefore considered a mere empirical tool which could give no analytical insight. The fact that their forecasts outperformed the former models, was attributed to misspecification problems in the structural equations econometric models. In 1974 Zellner and Palm established that time series models were a reduced form representation of structural econometric models and laid a bridge connecting them.

As a matter of fact, in many cases, it is impossible to estimate stochastic models in their structural form as we don't know all the relations and we do not have the relevant data. Therefore what we call structural model is, in the most of the cases, a reduced form whose specification is guided by our *a priori* information and the data available. Therefore, they can be seen as sparse multivariate time series models.

The standard or canonical multivariate ARIMA model is not in general sparse. It does not include contemporaneous dependence which is absorbed in the correlation structure of the multivariate residuals. Several authors have proposed methods of identifying or modifying these models by considering contemporaneous transformations of the series, see e.g. Box and Tiao (1977) and Tiao and Tsay (1989).

The resulting models are similar in appearance to the class of structural VAR models (see Hendry, 1995, p. 315) which include sparse contemporaneous dependence. In this thesis

we show how graphical modelling can assist us in identifying such sparse multivariate time series models.

In the first chapters of this thesis we give an overview of the relevant theory used in the following chapters. We begin by describing the basic facts about univariate time series, in chapter 2, ARMA models being derived as rational polynomial approximation of the Wold decomposition. Results about identification and estimation are presented for AR, MA and ARMA models.

In chapter 3, the arguments of the previous chapter are generalised to ARIMA models for multivariate series which are required for our later development. Two data sets, which will be used in later chapters, are introduced: three time series from the Italian monetary market and seven different terms to maturity of the U.S. dollar interest rate. This chapter includes a careful description, in terms of the interest rate example, of the efficient estimation of a multivariate ARMA(1,1) model. The results from this are used in later chapters.

In chapter 4 we introduce some basic definitions of graph theory used to present some results of graphical modelling. These results are then applied to the residuals from the multivariate ARMA(1,1) model previously fitted to the U.S. dollar interest rates. We show how a graphical model (GM) for these can be determined and estimated. These results provide a foundation for later application of GM to VAR and VARMA models.

In chapter 5 we begin to apply GM procedures to time series, in particular to univariate AR processes. We relate the partial correlation structure to a lemma by Ljung and Box which we use to determine the GM structure. We present an application to a simulated example of such models to evaluate the potential use of this approach.

The utility of GM in the univariate context is found to be limited, but in chapter 6 we demonstrate its value in the context of VAR processes. An important practical example is given in which a sparse structural VAR is determined for three series from the financial sector. Statistical evidence for different causal interpretations can be revealed and assessed by these methods.

In chapter 7 this is extended to the construction of a structural VARMA model for the multivariate series of term interest rates. The presence of MA parameters requires development of efficient methods of construction of the GM which builds on the estimation procedures described for VARMA models in chapter 3. This produces a much more

sparse model than the canonical VARMA and a model which reveals a very plausible causal dependence between the series.

We therefore believe that GM applied in this way provides a powerful empirical tool for the development of models for multiple time series which have a strong structural interpretation.

Chapter 2

Univariate Time Series Analysis

In this chapter we introduce some concepts of time series analysis which will be used in the following chapters.

2.1 Stochastic processes

We first give an informal definition of a stochastic process following presentations such as Cox and Miller (1965).

A *stochastic process* $\{X_t\}$ is a collection of random variables where the parameter t belongs to a parameter space (or index) T.

$$\{X_t\} = (\ldots X_1, X_2, \ldots, X_n, \ldots).$$

The set of possible values which each X_t can assume is commonly called the *state space*, S.

Other authors may use the term state space in a more specialised way meaning a Markovian process which is part of a state space model (Harvey, 1989, pp. 100-101).

The state space can be discrete or continuous, and can also be multidimensional. The parameter space T may also be discrete or continuous but we restrict ourselves to a single dimension which usually corresponds to time. Spatial processes in one or more dimensions may also be defined but are not considered in this thesis.

We give examples of four kind of process:

a) a process with discrete state space and discrete parameter space, e.g. the goals scored

weekly in the Italian football championship;
b) a process with continuous state space and discrete parameter space, e.g. the snow fall in Moscow every 24 hours;
c) a process with discrete state space and continuous parameter space, e.g. the number of Ph.D. students awaiting their viva;
d) a process with continuous state space and continuous parameter space, e.g. the electrocardiogram of a person registered for a certain period of time.

In this work we deal with processes of the second type where the parameter t is usually an integer which indexes a sequence of equally spaced points $\tau_t = ht$, where h is the time interval.

We distinguish between:

a) the whole process $\{X_t\}$;
b) the *sample function* $\{x_t\}$ which is a possible outcome or realisation of the process $\{X_t\}$, a sequence of real numbers in the state space;
c) the random variable X_t which is defined for a fixed value of the parameter t, e.g. the number of goals scored on a particular date;
d) the sample value x_t at a particular time, a single real number in S.

It is often convenient to relax this distinction in notation, to allow the context to indicate whether x_t is a sample value at a particular time or the whole process.

The following two examples describe two common classes of stochastic processes, the Markov chain and the martingale. In both case $T = \{0, 1, 2, \ldots\}$.

Example 2.1.1 *(Markov Chain). A stochastic process $\{X_t\}$ is called a Markov chain if*

$$P(X_t = s \mid X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) = P(X_t = s \mid X_{t-1} = x_{t-1})$$

for all $s, x_0, x_1, \ldots, x_t \in S$. □

Example 2.1.2 *(Martingale). A martingale is a stochastic process whose expected value at time t, conditional on the past, is the value of the process at the time t-1, that is*

$$E(X_{t+1} \mid X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = x_t \qquad \forall t$$

where x_0, x_1,...,x_t are the states at time 0,1,...t and have to be finite. A martingale difference is a process where

$$E(X_{t+1} \mid X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = 0$$

that is, the expected value of X at time $t-1$, given the information available at that time is equal to zero. □

2.2 Stationarity and Gaussianity

A stochastic process $\{X_t\} = (X_0, X_1, \ldots, X_n)$ is completely determined if we know the joint distributions of any finite selection of random variables belonging to the process. To achieve that for a general process is very difficult. Two assumptions are very useful, that the process belongs to the class of stationary and Gaussian processes.

2.2.1 Stationarity

Stationarity means the invariance of the probabilistic structure of a process over time. There are two different types of stationarity: strong (or strict) and weak (or second order). A process $\{X_t\}$ is strongly stationary if the families

$$X_k, X_{k+1}, \ldots, X_{k+n} \quad \text{and} \quad X_{k+h}, X_{k+1+h}, \ldots, X_{k+n+h}$$

have the same joint distribution for all k and n where $h > 0$. For second order stationarity of a process $\{X_t\}$ we just need that the mean and the variance of X are constant over time

$$\mathrm{E}(X_t) = \mu_x; \quad \mathrm{Var}(X_t) = \sigma_x^2 \qquad \forall t \tag{2.2.1}$$

and that the covariance between the processes $\{X_t\}$ and $\{X_{t-k}\}$ depends just on the lag k, i.e.:

$$\mathrm{Cov}(X_t, X_{t-k}) = \mathrm{E}\{(X_t - \mu_x)(X_{t-k} - \mu_x)\} = \gamma_k \qquad k \in \mathbb{Z}, \forall t. \tag{2.2.2}$$

Strong stationarity implies weak stationarity only if second order moment exists while weak stationarity does not necessarily imply strong stationarity.

Example 2.2.1 *(White noise). The simplest stationary process is a white noise or purely random process $\{\epsilon_t\}$, where all random variables have zero mean, constant variance and are uncorrelated, i.e.:*

$$E(\epsilon_t) = 0 \qquad \forall t; \tag{2.2.3}$$

$$E(\epsilon_t \epsilon_s) = \begin{cases} \gamma_0 = \sigma_\epsilon^2 & if\ t = s \\ \gamma_{t-s} = 0 & if\ t \neq s \end{cases} \qquad \forall t. \tag{2.2.4}$$

□

Example 2.2.2 *(The random walk). An example of a non stationary process is given by the random walk which in a simple form can be represented by the following equation*

$$X_t = X_{t-1} + \omega_t \tag{2.2.5}$$

when $\omega_1, \omega_2, \ldots, \omega_t$ are i.i.d. variables with mean μ_ω and variance σ_ω^2. Solving equation 2.2.5 in terms of ω_i, i=1,. . . ,n we have

$$X_t = (X_{t-2} + \omega_{t-1}) + \omega_t = \ldots = x_0 + \sum_{k=1}^{t} \omega_k$$

where x_0 is the fixed value of X_0. Hence the mean and the variance of the process will be

$$E(X_t) = x_0 + t\mu_\omega \qquad and \qquad Var(X_t) = t\sigma_\omega^2.$$

These depend upon time, for which reason we say that the process is not stationary. □

Weakly stationary processes can be parametrised by a simplified representation obtained with Wold's decomposition.

Proposition 2.2.3 *(Wold's decomposition). Any second-order stationary stochastic process $\{X_t\}$ with mean μ can be represented as the sum of two completely uncorrelated processes:*

$$X_t = Z_t + V_t \qquad \forall t \in T \tag{2.2.6}$$

where (see Cox and Miller, p. 288) Z_t is called the (purely) indeterministic and V_t is called the (purely) deterministic component. The deterministic component generally corresponds (Cox and Miller, pp. 314–315) to a small number of sinusoidal terms, which are however non ergodic (see section 2.4). It may be represented by

$$V_t = \mu + \sum_j [\alpha_j sin(\lambda_j t) + \beta_j cos(\lambda_j t)], \qquad 0 \leq \lambda_j \leq \pi \tag{2.2.7}$$

and

$$Z_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}; \qquad \psi_0 = 1, \quad \sum_{j=0}^{\infty} \psi_j^2 < +\infty \tag{2.2.8}$$

where $\{\epsilon_t\}$ is a white noise process characterised as the linear innovation of $\{X_t\}$ that is, $\epsilon_t = Z_t - \hat{Z}_{t,t-1}$ where $\hat{Z}_{t,t-1}$ is the minimum mean square error linear prediction of

Z_t in terms of $Z_{t-1}, Z_{t-2}, \ldots$. Therefore Z_t is a non deterministic component, because it cannot be exactly predicted from past values. The component V_t is called deterministic because it may be fitted as a time varying mean component with α_j and β_j treated as unknown constants. In fact, for any sample function $\{x_t\}$ of $\{X_t\}$ the coefficients α_j, β_j are sample values of zero mean random variables so that the population mean is constant.

In a single realisation $\{x_t\}$, modelling V_t as a deterministic component simplifies X_t by reducing the stochastic part to Z_t. □

A linear process is one for which the innovations ϵ_t form an independent series. That is ϵ_t is the error of the minimum variance prediction of X_t from all past values $X_{t-1}, X_{t-2}, \ldots$.

2.2.2 Gaussianity

Another useful property for the description of a process is Gaussianity. The process $\{X_t\}$ is Gaussian if for every k random variables belonging to the process their joint density function follows a multivariate normal density function

$$f(X_{t1}, X_{t2}, \ldots, X_{tk}) = (2\pi)^{-\frac{k}{2}} \mid \mathbf{\Sigma} \mid^{-\frac{1}{2}} exp\{-\frac{1}{2}(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\}, \qquad \mathbf{x} \in \mathbb{R}^k \quad (2.2.9)$$

where $\mathbf{x}$ is a vector of states, $\boldsymbol{\mu}$ is a vector of mean values and $\mathbf{\Sigma}$ is the covariance matrix of the k variables.

Hence a Gaussian process is determined if we know $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ whose generic element $\sigma_{r,s}$ is

$$\sigma_{r,s} = \mathrm{Cov}(X_r, X_s) = \mathrm{E}[(X_r - \mu_r)(X_s - \mu_s)], \qquad \forall r, s = 1, 2, \ldots, n.$$

If the process is second order stationary, μ will be constant for every t and the elements of $\mathbf{\Sigma}$ will depend only on the difference between the indices r and s which is invariant with respect to a translation over time. Hence for Gaussian processes weak stationarity and strong stationarity are coincident.

Mallows proved (1967), under not very restrictive assumptions, that linear processes are "nearly Gaussian" in the sense that many of the results which are true for the estimation of Gaussian processes are also true under the weaker assumption of a linear process.

2.3 Linear processes and invertibility

From equation 2.2.8 the non deterministic component of X_t has the representation in terms of the innovations ϵ_t:

$$Z_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}. \tag{2.3.1}$$

The coefficients ψ_j cannot have any set of values. They must satisfy the conditions:

$$\sum_{j=0}^{\infty} \psi_j^2 < \infty; \tag{2.3.2}$$

$$\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j \neq 0 \qquad \text{for} \qquad \mid z \mid < 1. \tag{2.3.3}$$

Equation 2.3.2 just states the finite variance of Z_t, $\sigma_z^2 = \sigma_\epsilon^2 \sum_{j=0}^{\infty} \psi_j^2$. Equation 2.3.3 (see par. 2.6.1) is known as the invertibility condition.

Provided these conditions hold, a process Z_t which is generated as the linear process 2.3.1 with ϵ_t independent, has as its innovations the sequence ϵ_t and coincides with its Wold's representation. Such a linear process is a useful first step in linear modelling of time series.

Example 2.3.1 *Let*

$$Z_t = \epsilon_t + 2\epsilon_{t-1}$$

then $\Psi(z) = (1 + 2z) = 0$ *for* $z = -\frac{1}{2}$. *Therefore the process is not invertible,* ϵ_t *is not the linear innovation of* Z_t *and the equation for* Z_t *is not the same as its Wold representation.* □

2.4 Time series and ergodicity

A time series x_t, $t = 1, \ldots, n$, is just a finite part of one realization of a stochastic process, that is a collection of sample values belonging to different random variables. Hence, to do inference about moment functions of the process from the observation of a single time series we need that the sample moments of an observed record of length n converge to the corresponding population moments as $n \to \infty$. This property is called ergodicity and for a particular stationary process it is defined with respect to the moments themselves. Formally, given a stationary process $\{X_t\}$, a moment function $E(\nu(X_t))$ and its temporal

estimator $\hat{\nu}_t(x_t)$, the process is ergodic with respect to the moment if

$$\lim_{n\to\infty} \hat{\nu}_n(x_t) = \lim_{n\to+\infty} \frac{1}{n}\sum_{t=1}^{n} \nu(x_t) = \int_{-\infty}^{+\infty} \nu(x_t) f(x_t) dx_t = E(\nu(x_t)).$$

In practice it represents for processes what the law of large numbers is for independent random variables.

Stationarity is necessary for ergodicity but the two concepts are not coincident.

A detailed exposition of the condition for ergodicity can be found in Doob (1953). In the rest of this thesis we shall assume ergodicity for the first two moments of the time series we analyse.

2.5 Autocorrelation functions

There are two functions which are very useful in order to identify a time series model: the *autocorrelation function* (ACF) and the *partial autocorrelation function* (PACF). We shall describe also a generalisation of the ACF which is called the *autocorrelation generating function* (ACGF).

2.5.1 The autocorrelation function

For a stochastic process $\{X_t\}$, the autocovariance function, $\gamma(k)$, is given by

$$\gamma(k) = \mathrm{Cov}(X_t, X_{t-k}). \tag{2.5.1}$$

Correspondingly the autocorrelation function (ACF), $\rho(k)$, is defined as

$$\rho(k) = \mathrm{Corr}(X_t, X_{t-k}) = \frac{\mathrm{Cov}(X_t, X_{t-k})}{\sqrt{\mathrm{Var}(X_t)\mathrm{Var}(X_{t-k})}}. \tag{2.5.2}$$

Of course, it is meaningful to take this function into consideration only for stationary processes. In this case because $\mathrm{Var}(X_t)$=$\mathrm{Var}(X_{t-k})$, $\forall t, k$, 2.5.2 becomes

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}. \tag{2.5.3}$$

The autocorrelation function has several properties:
a) $\rho(0)$=1;

b) it is an even function. In fact $\gamma(k) = \text{Cov}(X_t, X_{t-k}) = \text{Cov}(X_{t-k}, X_t) = \text{Cov}(X_t, X_{t+k}) = \gamma(-k)$, hence

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\gamma(-k)}{\gamma(0)} = \rho(-k).$$

c) $\mid \rho(k) \mid \leq 1$. In fact because of the Cauchy-Schwartz inequality

$$(\text{E}[(X_t - \text{E}(X_t))(X_{t-k} - \text{E}(X_{t-k}))])^2 \leq \text{E}[(X_t - \text{E}(X_t))^2]\text{E}[(X_{t-k} - \text{E}(X_{t-k}))^2] \quad (2.5.4)$$

and because of the stationarity of the process

$$\text{E}[(X_t - \text{E}(X_t))^2] = \text{Var}\,(X_t) = \text{Var}\,(X_{t-k}) = \text{E}[(X_{t-k} - \text{E}(X_{t-k}))^2].$$

Hence 2.5.4 becomes $\gamma^2(k) \leq \gamma^2(0)$ which implies

$$\mid \gamma(k) \mid \leq \mid \gamma(0) \mid \Leftrightarrow \frac{\mid \gamma(k) \mid}{\mid \gamma(0) \mid} \leq \frac{\mid \gamma(0) \mid}{\mid \gamma(0) \mid} \Leftrightarrow \mid \rho(k) \mid \leq 1;$$

d) the ACF is invariant to any linear transformation of the process of the form $X_t \rightarrow a + bX_t$;

e) the Toeplitz matrix of order k related to the ACF function

$$\mathbf{P}_{(k)} = \begin{bmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(k-1) \\ \rho(1) & 1 & \rho(1) & \dots & \rho(k-2) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \rho(k-1) & \rho(k-2) & \dots & \dots & 1 \end{bmatrix} \quad (2.5.5)$$

is positive semidefinite;

f) if the ACF ρ(k) of a stationary process satisfies the condition $\sum \mid \rho(k) \mid < \infty$, then there is a function $\eta(\omega)$, $-\pi \leq \omega \leq \pi$, such that

$$\eta(\omega) \geq 0; \qquad \int_{-\pi}^{\pi} \eta(\omega) d\omega = 1; \qquad \int_{-\pi}^{\pi} \eta(\omega)\cos(\omega k)\ d\omega = \rho(k)$$

and

$$\eta(\omega) = \frac{1}{2\pi}\left[1 + 2\sum_{k=1}^{\infty} \rho(k)\cos(\omega k)\right]. \quad (2.5.6)$$

The function $\eta(\omega)$ is called *spectral density function.*

The autocovariance of X_t may therefore be given by

$$\gamma(k) = \int_{-\pi}^{\pi} \sigma_x^2 \eta(\omega)\cos(\omega k)\ d\omega = \int_{-\pi}^{\pi} S(\omega)\cos(\omega k)\ d\omega \quad (2.5.7)$$

where

$$S(\omega) = \sigma_x^2 \eta(\omega) = \frac{1}{2\pi}\left[\gamma(0) + 2\sum_{k=1}^{\infty} \gamma(k)\cos(\omega k)\right]$$

is called the *spectrum* of X_t.

It is commonly assumed for practical time series analysis that the non-deterministic part Z_t of a stationary process (as represented in 2.2.6) is a linear process with continuous spectral density function $\eta(\omega)$.

2.5.2 The autocorrelation generating function

The *autocovariance generating function* is a useful transformation of the autocovariance function and it is given by

$$G(B) = \sum_{k=-\infty}^{\infty} \gamma(k)B^k = \gamma(0) + \sum_{k=1}^{\infty} \gamma(k)[B^k + B^{-k}] \tag{2.5.8}$$

because $\gamma(k) = \gamma(-k)$ and then the *autocorrelation generating function* (ACGF) is defined as

$$\wp(B) = \sum_{k=-\infty}^{\infty} \rho(k)B^k = 1 + \sum_{k=1}^{\infty} \rho(k)[B^k + B^{-k}] = \frac{G(B)}{\gamma(0)}. \tag{2.5.9}$$

It is seen shortly why we use B for the dummy variable rather than, for example, the variable Z which is normally used in generating functions.

Example 2.5.1 *Consider the white noise process* $\{\epsilon_t\}$. *It has* $\gamma(0) = \sigma^2$ *and* $\gamma(k) = 0$ *for* $k \neq 0$, *hence*

$$G(B) = \sum_{k=-\infty}^{\infty} \gamma(k)B^k = \sigma^2 \qquad \textit{and} \qquad \wp(B) = \frac{G(B)}{\gamma(0)} = 1. \tag{2.5.10}$$

□

If we consider

$$Y_t = \sum_{i=0}^{\infty} \alpha_i X_{t-i} = \alpha_0 X_t + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots, \tag{2.5.11}$$

using the backward shift operator B defined as

$$\mathrm{B}^n \epsilon_t = \epsilon_{t-n},$$

we can write

$$Y_t = \alpha_0 + \alpha_1 \mathrm{B} + \alpha_2 \mathrm{B}^2 + \ldots = \alpha(\mathrm{B}) \tag{2.5.12}$$

then

$$G_y(\mathrm{B}) = \alpha(\mathrm{B})\alpha(\mathrm{B}^{-1})G_x(\mathrm{B}) = A(\mathrm{B})G_x(\mathrm{B}) \tag{2.5.13}$$

where $A(\mathrm{B}) = \alpha(\mathrm{B})\alpha(\mathrm{B}^{-1})$, considering B as acting upon the index h, that is

$$\gamma_y(h) = \alpha(\mathrm{B})\alpha(\mathrm{B}^{-1})\gamma_x(h) \tag{2.5.14}$$

obtaining in this way a linear filter relating the covariance sequences.

Derivation

$$\begin{aligned}
\gamma_y(k) &= \mathrm{Cov}\,(Y_t, Y_{t-k}) \\
&= \mathrm{Cov}\left(\sum_i \alpha_i X_{t-i}, \sum_j \alpha_j X_{t-j}\right) \\
&= \sum_i \sum_j \alpha_i \alpha_j \mathrm{Cov}\,(X_{t-i} X_{t-j}) \\
&= \sum_i \sum_j \alpha_i \alpha_j \gamma_x(k+i-j) \qquad (2.5.15)
\end{aligned}$$

hence

$$\begin{aligned}
G_y(\mathrm{B}) &= \sum_k \gamma_y(k)\mathrm{B}^k \\
&= \sum_k \sum_i \sum_j \mathrm{B}^k \alpha_i \alpha_j \gamma_x(k+i-j) \\
&= \sum_k \sum_i \sum_j \alpha_i \mathrm{B}^{-i} \alpha_j \mathrm{B}^j \mathrm{B}^{k+i-j} \gamma_x(k+i-j).
\end{aligned}$$

Let $h = k + i - j$, hence

$$\begin{aligned}
G_y(\mathrm{B}) &= \sum_h \sum_i \sum_j \alpha_i \mathrm{B}^{-i} \alpha_j \mathrm{B}^j \mathrm{B}^h \gamma_x(h) \\
&= \alpha(\mathrm{B})\alpha(\mathrm{B}^{-1}) G_x(\mathrm{B}) \qquad (2.5.16) \\
&= A(\mathrm{B}) G_x(\mathrm{B}). \qquad (2.5.17)
\end{aligned}$$

□

Example 2.5.2 *Suppose* $Y_t = \alpha_0 + \alpha_1 X_{t-1}$, *then* $\alpha(B) = (\alpha_0 + \alpha_1 B)$, *hence*

$$\begin{aligned}
A(B) &= (\alpha_0 + \alpha_1 B)(\alpha_0 + \alpha_1 B^{-1}) \\
&= \alpha_0^2 + \alpha_0\alpha_1 B^{-1} + \alpha_0\alpha_1 B + \alpha_1^2 \\
&= (\alpha_0^2 + \alpha_1^2) + \alpha_0\alpha_1 B + \alpha_0\alpha_1 B^{-1}
\end{aligned}$$

and

$$\gamma_y(k) = (\alpha_0^2 + \alpha_1^2)\gamma_x(k) + \alpha_0\alpha_1\gamma_x(k-1) + \alpha_0\alpha_1\gamma_x(k+1). \ \square$$

If $X \sim WN(0, \sigma_x^2)$ then

$$G_y(\mathrm{B}) = \alpha(\mathrm{B})\alpha(\mathrm{B}^{-1})\sigma_x^2 \tag{2.5.18}$$

and similarly for the stochastic part of the Wold's decomposition

$$G_z(\mathrm{B}) = \psi(\mathrm{B})\psi(\mathrm{B}^{-1})\sigma_\epsilon^2. \tag{2.5.19}$$

Substituting $e^{i\omega}$ to B in the ACGF obtaining the spectrum

$$\begin{aligned} S(\omega) &= \frac{1}{2\pi} G(e^{i\omega}) \\ &= \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j(k) e^{i\omega k} \\ &= \frac{1}{2\pi} \left[\gamma_0 + 2 \sum_{j=1}^{\infty} \gamma_j(k) \cos(\omega k) \right] \end{aligned} \tag{2.5.20}$$

thus from 2.5.16

$$\begin{aligned} S_y(\omega) &= \alpha(e^{i\omega})\alpha(e^{-i\omega}) S_x(\omega) \\ &= \left| \alpha(e^{i\omega}) \right|^2 S_x(\omega) \end{aligned} \tag{2.5.21}$$

which leads, when $X \sim WN(0, \sigma^2)$, to

$$S_z(e^{i\omega}) = \left| \Psi(e^{i\omega}) \right|^2 \sigma_\epsilon^2. \tag{2.5.22}$$

Equation 2.5.18 reveals the *time reversibility* of linear processes. In fact $Z_t = \Phi(\mathrm{B})a_t$ has the same ACGF as $W_t = \Phi(\mathrm{B}^{-1})a_t$:

$$G_Z(\mathrm{B}) = \sigma^2 \Phi(\mathrm{B})\Phi(\mathrm{B}^{-1}) = \sigma^2 \Phi(\mathrm{B}^{-1})\Phi(\mathrm{B}) = G_W(\mathrm{B})$$

thus sequences $\{z_t\}$ and $\{z_{n-t-1}\}$, $t = 1, 2, \ldots, n$ have the same stochastic structure when considering moments of first and second order. This result motivates the back-forecasting, which is used to predict past values from more recent values of the time series.

Using the concept of back-forecasting and the autocovariance structure it is possible to derive an apparently strange result.

Proposition 2.5.3 *There exists a stationary process, Z_t for which*

$$Z_t = \frac{1}{\phi} Z_{t-1} + \epsilon_t \qquad , 0 < \mid \phi \mid < 1 \tag{2.5.23}$$

with stationary autocovariance function $\gamma(k) = \phi^k$, where ϵ_t is white noise.

Remark: Such a process cannot be generated from ϵ_t in the sequence t=0,1,2,... because the calculations are unstable.

Derivation

Consider the model

$$Z_t = \phi Z_{t-1} + e_t \tag{2.5.24}$$

where $\mid \phi \mid < 1$, $e_t \sim$WN and $e_t \perp Z_{t-1}$,

$$\begin{aligned} Z_t &= \phi(\phi Z_{t-2} + e_{t-1}) + e_t \\ &= \phi^2(\phi Z_{t-3} + e_{t-2}) + \phi e_{t-1} + e_t \\ &= e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \dots \end{aligned} \tag{2.5.25}$$

hence for this process

$$\gamma(k) = \text{Cov}\,(Z_t, Z_{t-k}) = \phi^k \sigma_Z^2. \tag{2.5.26}$$

Now consider the time reversal model for Z_t

$$Z_t = \phi Z_{t+1} + f_t \tag{2.5.27}$$

where ϕ is the same parameter as in 2.5.24, $f_t \sim$WN and $f_t \perp Z_{t+1}$, then rearranging

$$Z_{t+1} = \frac{1}{\phi} Z_t - \frac{1}{\phi} f_t \tag{2.5.28}$$

let $\epsilon_{t+1} = -\frac{1}{\phi} f_t$ and rearrange then

$$Z_t = \frac{1}{\phi} Z_{t-1} + \epsilon_t \tag{2.5.29}$$

where

$$\epsilon_t = Z_t - \frac{1}{\phi} Z_{t-1}. \tag{2.5.30}$$

We can verify that ϵ_t is still a white noise process, in fact

$$\begin{aligned} \text{Cov}\,(\epsilon_t, \epsilon_{t-k}) &= \text{Cov}\,(Z_t - \frac{1}{\phi} Z_{t-1}, Z_{t-k} - \frac{1}{\phi} Z_{t-k-1}) \\ &= \gamma(k) - \frac{1}{\phi}\gamma(k+1) - \frac{1}{\phi}\gamma(k-1) + \frac{1}{\phi^2}\gamma(k). \end{aligned} \tag{2.5.31}$$

As said before model 2.5.24 and model 2.5.27 have the same covariance structure, hence we can substitute value 2.5.26 to the autocovariances in 2.5.31, thus

$$\text{Cov}\,(\epsilon_t, \epsilon_{t-k}) = \sigma_Z^2(\phi^k - \frac{1}{\phi}\phi^{k+1} - \frac{1}{\phi}\phi^{k-1} + \frac{1}{\phi^2}\phi^k) = 0. \tag{2.5.32}$$

Because $\epsilon_t = -\frac{1}{\phi} f_{t-1}$ is correlated with Z_{t-1}, it is clear that although ϵ_t is WN, it is not the (forward) linear innovation of Z_t in 2.5.23.

2.5.3 The partial autocorrelation function

The ACF $\rho(k)$ describes the correlation between the value of a process at times t and $t \pm k$ without referring to values at intermediate lags. The partial autocorrelation function, π(k), describes the same correlation conditional upon the values at intermediate times,

$$\pi(k) = \mathrm{Corr}(Z_t, Z_{t-k} \mid Z_{t-1}, \ldots, Z_{t-k+1})) \tag{2.5.33}$$

(see Whittaker, 1990, pp. 134–137).

Conditional autocorrelation can be obtained by considering a basic result of regression analysis (see Huang, 1970, p.13): in a regression like

$$Y = \phi X + \epsilon \tag{2.5.34}$$

the regression parameter ϕ is given by

$$\phi = \rho \frac{\sigma_y}{\sigma_x} \tag{2.5.35}$$

where ρ is the correlation index between X and Y and σ_y and σ_x are respectively the mean square errors of Y and X. If we consider the regression

$$Z_t = \phi Z_{t-1} + \epsilon_t$$

then, because of stationarity $\sigma_{x_t} = \sigma_{x_{t-1}}$ and $\phi = \rho$. Such a concept can be extended to consideration of the last parameter regressions like

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_m Z_{t-m} + \ldots + \phi_p Z_{t-p} + \epsilon_t.$$

In fact the partial correlation coefficient, $\pi(p)$ is always proportional to the actual coefficient of the final variable, ϕ_p, in a regression, the relationship being

$$\phi_p = \pi(p) \left(\frac{\mathrm{Var}\left(Z_t \mid Z_{t-1}, \ldots, Z_{t-p+1}\right)}{\mathrm{Var}\left(Z_{t-p} \mid Z_{t-1}, \ldots, Z_{t-p+1}\right)} \right)^{\frac{1}{2}}. \tag{2.5.36}$$

Because of the time reversible property of stationary time series the variances in the ratio are identical and $\phi_p = \pi(p)$. Hence it is apparent that estimating ϕ_i, $i = 0, 1, 2, \ldots$, in regressions where the maximum lag p is equal to i we obtain an estimate of $\pi(i)$. A simple way to estimate ϕ_i is given by the *Yule-Walker equations*.

We are interested in predicting Z_t from $Z_{t-1}, \ldots, Z_{t-p}$ by a linear equation of the form $Z_t = \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + \epsilon_t$. The coefficients are characterised by the property that the error ϵ_t is uncorrelated with $Z_{t-1}, \ldots, Z_{t-p}$, writing

$$Z_t - \phi_1 Z_{t-1} - \phi_2 Z_{t-2} - \ldots - \phi_p Z_{t-p} = a_t \tag{2.5.37}$$

multiplying both sides by Z_{t-k} we obtain

$$Z_t Z_{t-k} - \phi_1 Z_{t-1} Z_{t-k} - \phi_2 Z_{t-2} Z_{t-k} - \ldots - \phi_p Z_{t-p} Z_{t-k} = a_t Z_{t-k} \tag{2.5.38}$$

and considering that

$$\mathrm{E}(a_t Z_{t-k}) = \begin{cases} \sigma_a^2 & , k = 0 \\ 0 & , k \neq 0 \end{cases} \tag{2.5.39}$$

and assuming that Z_t is mean corrected

$$\mathrm{E}[Z_{t-j} Z_{t-k}] = \gamma(k-j) \qquad , \forall j, k. \tag{2.5.40}$$

If we take the expectation of 2.5.38 we obtain

$$\gamma(k) = \phi_1 \gamma(k-1) + \phi_2 \gamma(k-2) + \ldots + \phi_p \gamma(k-p) \qquad , k > 0 \tag{2.5.41}$$

hence for $k = 0, 1, 2, \ldots$

$$\begin{cases} \gamma(0) = \phi_1 \gamma(1) + \phi_2 \gamma(2) + \phi_3 \gamma(3) + \ldots + \phi_p \gamma(p) + \sigma_a^2 \\ \gamma(1) = \phi_1 \gamma(0) + \phi_2 \gamma(1) + \phi_3 \gamma(2) + \ldots + \phi_p \gamma(p-1) \\ \gamma(2) = \phi_1 \gamma(1) + \phi_2 \gamma(0) + \phi_3 \gamma(1) + \ldots + \phi_p \gamma(p-2) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases} \tag{2.5.42}$$

The first of these gives

$$\frac{\sigma_a^2}{\sigma_Z^2} = (1 - \phi_1 \rho(1) - \ldots - \phi_p \rho(p).$$

Taking these equations for $\gamma(1), \ldots, \gamma(p)$ gives the Yule-Walker equations. If we divide $\gamma(1), \gamma(2), \ldots$ by $\gamma(0)$ we have

$$\begin{cases} \rho(1) = \phi_1 + \phi_2 \rho_(1) + \phi_3 \rho(2) + \ldots + \phi_p \rho(p-1) \\ \rho(2) = \phi_1 \rho(1) + \phi_2 + \phi_3 \rho(1) + \ldots + \phi_p \rho(p-2) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \rho(p) = \phi_1 \rho(p-1) + \phi_2 \rho(p-2) + \phi_3 \rho(p-3) + \ldots + \phi_p \end{cases} \tag{2.5.43}$$

System 2.5.3 can be represented in matrix form as

$$\boldsymbol{\rho} = \boldsymbol{P}_{(p)} \boldsymbol{\phi}$$

so that

$$\boldsymbol{\phi} = \boldsymbol{P}_{(p)}^{-1} \boldsymbol{\rho}. \tag{2.5.44}$$

Hence, estimating the last parameter ϕ_p in systems with $p = 1, 2, \ldots$ we obtain $\pi(1), \pi(2), \ldots$. Applying Cramer's rule we have

$$\pi(k) = \frac{\mid \mathbf{Q}_{(k)} \mid}{\mid \mathbf{P}_{(k)} \mid}, \qquad k = 1, 2, \ldots \tag{2.5.45}$$

where $\mathbf{P}_{(k)}$ is the Toeplitz matrix of order k and $\mathbf{Q}_{(k)}$ is a matrix which has only the last column different from $\mathbf{P}_{(k)}$ (see Box and Jenkins, 1976, p. 64),

$$\mathbf{Q}_{(k)} = \begin{bmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(k-2) & \rho(1) \\ \rho(1) & 1 & \rho(1) & \dots & \rho(k-3) & \rho(2) \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ \rho(k-1) & \rho(k-2) & \dots & \dots & \rho(1) & \rho(k) \end{bmatrix}. \tag{2.5.46}$$

Conventionally, $\pi(0) = 1$ and it is easy to compute $\pi(1)$, $\pi(2)$ and $\pi(3)$.

Example 2.5.4

$$\pi(1) = \frac{\mid \mathbf{Q}_{(1)} \mid}{\mid \mathbf{P}_{(1)} \mid} = \frac{\rho(1)}{\rho(0)} = \rho(1).$$

□

Example 2.5.5

$$\pi(2) = \frac{\mid \mathbf{Q}_{(2)} \mid}{\mid \mathbf{P}_{(2)} \mid} = \frac{\begin{vmatrix} 1 & \rho(1) \\ \rho(1) & \rho(2) \end{vmatrix}}{\begin{vmatrix} 1 & \rho(1) \\ \rho(1) & 1 \end{vmatrix}} = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2}.$$

□

Example 2.5.6

$$\pi(3) = \frac{\mid \mathbf{Q}_{(3)} \mid}{\mid \mathbf{P}_{(3)} \mid} = \frac{\begin{vmatrix} 1 & \rho(1) & \rho(1) \\ \rho(1) & 1 & \rho(2) \\ \rho(2) & \rho(1) & \rho(3) \end{vmatrix}}{\begin{vmatrix} 1 & \rho(1) & \rho(2) \\ \rho(1) & 1 & \rho(1) \\ \rho(2) & \rho(1) & 1 \end{vmatrix}} = \frac{\rho(3)[1 - \rho^2(1)] + \rho(1)[\rho^2(1) + \rho^2(2) - 2\rho(2)]}{[1 - \rho^2(2)][1 + \rho(2) - 2\rho^2(1)]}.$$

□

A recursive method to estimate the partial autocorrelation function from the correlation function is given by the *Durbin-Levinson algorithm*. To do this we must distinguish the fact that the coefficients depend on the number k of lagged values used in this prediction so we then write e.g. $\phi_{j,k}$ for the j^{th} coefficient in the prediction using k lagged values.

From 2.5.36 we know that

$$\pi(1) = \rho(1) \tag{2.5.47}$$

where $\rho(k)$ is the autocorrelation coefficient of order k in the autoregressive model of order i.

Now consider the first lag regression

$$Z_t = \phi_{1,1} Z_{t-1} + e_{t,1} \tag{2.5.48}$$

where $e_{t,1}$ is the prediction error of Z_t from Z_{t-1}, that is

$$e_{t,1} = Z_t - \phi_{1,1} Z_{t-1} \tag{2.5.49}$$

with

$$\text{Var}\,(e_{t,1}) = \sigma_1^2 = \sigma_z^2(1 - \phi_{1,1}^2) = \sigma_z^2(1 - \pi(1)^2) \tag{2.5.50}$$

and consider also the prediction error $f_{t-2,1}$ in the time-reversal model of Z_{t-2} from Z_{t-1}, that is

$$f_{t-2,1} = Z_{t-2} - \phi_{1,1} Z_{t-1} \tag{2.5.51}$$

with $\text{Var}\,(f_{t-2,1}) = \sigma_1^2$. From 2.5.49 and 2.5.51 we have

$$\begin{aligned} \pi(2) &= \text{Corr}\,(Z_t, Z_{t-2} \mid Z_{t-1}) \\ &= \text{Corr}\,(e_{t,1}, f_{t-2,1}). \end{aligned} \tag{2.5.52}$$

On the other hand

$$\text{Cov}\,(e_{t,1}, f_{t-2,1}) = \text{Cov}\,(e_{t,1}, Z_{t-2} - \phi_{1,1} Z_{t-1}). \tag{2.5.53}$$

From equation 2.5.49 we know that $e_{t,1}$ is uncorrelated with Z_{t-1} hence

$$\begin{aligned} \text{Cov}\,(e_{t,1}, f_{t-2,1}) &= \text{Cov}\,(e_{t,1}, Z_{t-2}) \\ &= \text{Cov}\,(Z_t - \phi_{1,1} Z_{t-1}, Z_{t-2}) \\ &= \sigma_z^2(\rho(2) - \phi_{1,1}\rho(1)) \end{aligned} \tag{2.5.54}$$

and because $\text{Var}\,(e_{t,1}) = \sigma_1^2 = \text{Var}\,(f_{t-2,1})$, we have

$$\text{Corr}\,(e_{t,1}, f_{t-2,1}) = \pi(2) = \frac{\sigma_z^2(\rho(2) - \phi_{1,1}\rho(1))}{\sigma_1^2}. \tag{2.5.55}$$

Substituting for σ_1^2 we have

$$\begin{aligned} \text{Corr}\,(e_{t,1}, f_{t-2,1}) = \pi(2) &= \frac{\sigma_z^2(\rho(2) - \phi_{1,1}\rho(1))}{\sigma_z^2(1 - \pi(1)^2)} \\ &= \frac{\rho(2) - \phi_{1,1}\rho(1)}{1 - \pi(1)^2} \\ &= \frac{\rho(2) - \phi_{1,1}\rho(1)}{1 - \rho_1^2}. \end{aligned} \tag{2.5.56}$$

Furthermore, the new prediction error variance is

$$\sigma_2^2 = \text{Var}\,(e_{1,t})(1-\pi(2)^2) = \sigma_1^2(1-\pi(2)^2) = \sigma_z^2(1-\pi(1)^2)(1-\pi(2)^2). \tag{2.5.57}$$

Similarly to obtain $\pi(3)$ we need to consider the equations of the prediction errors from a second lag regression

$$e_{t-2} = Z_t - \phi_{2,1}Z_{t-1} - \phi_{2,2}Z_{t-2} \tag{2.5.58}$$

and

$$f_{t-3,2} = Z_{t-3} - \phi_{2,1}Z_{t-2} - \phi_{2,2}Z_{t-1} \tag{2.5.59}$$

then

$$\text{Cov}\,(e_{t,2}, f_{t-3,2}) = \text{Cov}\,(e_{t,2}, Z_{t-3} - \phi_{2,1}Z_{t-2} - \phi_{2,2}Z_{t-1}) \tag{2.5.60}$$

from equation 2.5.58 we know that $e_{t,2}$ is uncorrelated with Z_{t-1} and Z_{t-2} hence

$$\begin{aligned} \text{Cov}\,(e_{t,2}, f_{t-3,2}) &= \text{Cov}\,(e_{t,2}, Z_{t-3}) \\ &= \text{Cov}\,(Z_t - \phi_{2,1}Z_{t-1} - \phi_{2,2}Z_{t-2}, Z_{t-3}) \\ &= \sigma_z^2(\rho(3) - \phi_{2,1}\rho(2) - \phi_{2,2}\rho(1)) \end{aligned} \tag{2.5.61}$$

and because $\text{Var}\,(e_{t,2}) = \sigma_2^2 = \text{Var}\,(f_{t-3,2})$, we have

$$\begin{aligned} \text{Corr}\,(e_{t,2}, f_{t-3,2}) = \pi(3) &= \frac{\sigma_z^2(\rho(3) - \phi_{2,1}\rho(2) - \phi_{2,2}\rho(1))}{\sigma_2^2} \\ &= \frac{\rho(3) - \phi_{2,1}\rho(2) - \phi_{2,2}\rho(1)}{[1-\pi(1)^2][1-\pi(2)^2]} \end{aligned} \tag{2.5.62}$$

where substituting the values for $\pi(1)$ and $\pi(2)$ we obtain $\pi(3)$ as a function of $\rho(1)$, $\rho(2)$ and $\rho(3)$.

In general once we have obtained values for $\pi(1), \ldots \pi(k)$ as functions of $\rho(1), \ldots, \rho(k)$ it is possible to express

$$\text{Corr}\,(e_{t,k}, f_{t-k-1,k}) = \pi(k+1) = \frac{\rho(k+1) - \phi_{k,1}\rho(k) - \ldots - \phi_{k,k}\rho(1)}{[1-\pi(1)^2][1-\pi(2)^2]\ldots[1-\pi(k)^2]} \tag{2.5.63}$$

in terms of $\rho(1), \ldots, \rho(k+1)$. See Newton, 1988, p.88 for further explanation of this algorithm. It has the advantage of computing the coefficients of lagged prediction of increasing order in a very efficient manner.

2.6 ARIMA models

From numerical analysis we know that we can approximate a continuous function over a finite interval by a polynomial of finite degree,

$$g_m(x) = a_0 + a_1x + a_2x^2 + \ldots + a_mx^m. \tag{2.6.1}$$

We can apply this to approximating a continuous spectrum S(ω) by a polynomial in $c = cos\omega$, i.e.

$$S(\omega) \approx a_0 + a_1 c + a_2 c^2 + \ldots + a_m c^m. \tag{2.6.2}$$

Such a polynomial may also be written as

$$S(\omega) \approx A_0 + A_1 \mathrm{cos}\omega + A_2 \mathrm{cos}2\omega + \ldots + A_m \mathrm{cos}m\omega \tag{2.6.3}$$

because cos(ωk) is the Chebychev polynomial of degree k in c. This is equivalent to approximating the autocovariances of the process by

$$\frac{1}{\pi}\gamma_k = \begin{cases} A_k, & k = 1, \ldots, m \\ 0, & k > m \end{cases} .$$

Other approximations used in numerical analysis are the reciprocal of polynomials

$$S(\omega) \approx (B_0 + B_1 \mathrm{cos}\omega + B_2 \mathrm{cos}2\omega + \ldots + B_n \mathrm{cos}\omega)^{-1} \tag{2.6.4}$$

or rational functions

$$S(\omega) \approx \frac{(A_0 + A_1 \mathrm{cos}\omega + A_2 \mathrm{cos}2\omega + \ldots + A_m \mathrm{cos}m\omega)}{(B_0 + B_1 \mathrm{cos}\omega + B_2 \mathrm{cos}2\omega + \ldots + B_n \mathrm{cos}n\omega)} \tag{2.6.5}$$

where B_0 is set to 1 to ensure a unique ratio.

The class of rational functions contains the other two classes and with a given number of coefficients is a wider class than that of the polynomials with the same number of coefficients and so has a greater capacity of approximation. In particular it can better approximate functions with sharp peaks or rapid change in gradient.

For example, the class of rational functions with 3 parameters includes

$$A_0 + A_1 \mathrm{cos}\omega + A_2 \mathrm{cos}2\omega; \qquad \frac{1}{B_0 + B_1 \mathrm{cos}\omega + B_2 \mathrm{cos}2\omega}; \qquad \frac{A_0 + A_1 \mathrm{cos}\omega}{1 + B_1 \mathrm{cos}\omega}.$$

Consider now a process represented by the general linear model:

$$X_t = \psi_0 \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + \ldots$$

that is

$$X_t = (\psi_0 + \psi_1 \mathrm{B} + \psi_2 \mathrm{B}^2 + \ldots)\epsilon_t = \Psi(\mathrm{B})\epsilon_t \tag{2.6.6}$$

in terms of the function Ψ(z) defined in 2.3.3. The spectrum of X_t is then given by

$$S(\omega) = \sigma_\epsilon^2 \mid \Psi(e^{i\omega}) \mid^2 . \tag{2.6.7}$$

If S(ω) is a rational function of the form 2.6.5 then it may be shown that $\Psi(z)$ is also a rational function in z

$$\Psi(z) = \frac{1 - \theta z - \ldots - \theta_q z^q}{1 - \phi z - \ldots - \phi_p z^p} \tag{2.6.8}$$

and using the back-shift notation the process X_t is then represented as

$$X_t = \frac{1 - \theta \mathrm{B} - \ldots - \theta_q \mathrm{B}^q}{1 - \phi \mathrm{B} - \ldots - \phi_p \mathrm{B}^p} \epsilon_t. \tag{2.6.9}$$

Example 2.6.1 *Consider*

$$\Psi(B) = \frac{1 - \theta B}{1 - \phi B}$$

$$\begin{aligned} S(\omega) &= \sigma^2 \left| \frac{1 - \theta e^{i\omega}}{1 - \phi e^{i\omega}} \right|^2 \\ &= \sigma^2 \frac{(1 - \theta e^{i\omega})(1 - \theta e^{-i\omega})}{(1 - \phi e^{i\omega})(1 - \phi e^{-i\omega})} \\ \textit{using } e^{i\omega} + e^{-i\omega} &= 2cos\omega \qquad (2.6.10) \\ &= \sigma^2 \frac{[(1 - \theta^2) - 2\theta cos\omega]}{[(1 - \phi^2) - 2\phi cos\omega]} \\ &= \frac{A_0 + A_1 cos\omega}{B_0 + B_1 cos\omega}. \qquad (2.6.11) \end{aligned}$$

□

This development motivates ARMA models, by writing 2.6.9 in the form

$$(1 - \phi \mathrm{B} - \ldots - \phi_p \mathrm{B}^p) X_t = (1 - \theta \mathrm{B} - \ldots - \theta_q \mathrm{B}^q) \epsilon_t \tag{2.6.12}$$

or

$$X_t = \phi X_{t-1} + \ldots + \phi_p X_{t-p} + \epsilon_t - \theta \epsilon_{t-1} - \ldots - \theta_q \epsilon_{t-q}. \tag{2.6.13}$$

2.6.1 Moving average models

This corresponds to the polynomial 2.6.3 of degree q approximating the spectrum, that is

$$Z_t = \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + \ldots + \psi_q \epsilon_{t-q} \tag{2.6.14}$$

where q is lower or equal to the degree of 2.6.3. Conventionally 2.6.14 is represented as

$$\begin{aligned} Z_t &= \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q} = \\ &= (1 - \theta_1 \mathrm{B} - \theta_2 \mathrm{B}^2 - \ldots - \theta_q \mathrm{B}^q) \epsilon_t = \Theta(\mathrm{B}) \epsilon_t \end{aligned} \tag{2.6.15}$$

which is called moving average model of order q and it is indicated as MA(q).

Example 2.6.2 *A MA(1) model is given by*

$$Z_t = \epsilon_t + \psi_1 \epsilon_{t-1} = (1 - \theta B)\epsilon_t. \tag{2.6.16}$$

□

We may be interested in inverting the relation between ϵ_t and Z_t, that is obtaining the time series of innovations, ϵ, from present and past observations of Z. Nevertheless to do this we need some conditions. In fact 2.6.7 may be represented as

$$\epsilon_t = \frac{1}{(1 - \theta_1 \mathrm{B} - \theta_2 \mathrm{B}^2 - \ldots - \theta_q \mathrm{B}^q)} Z_t \tag{2.6.17}$$

and factorising

$$\epsilon_t = \frac{1}{\alpha(r_1 - \mathrm{B})(r_2 - \mathrm{B}) \ldots (r_q - \mathrm{B})} Z_t \tag{2.6.18}$$

where r_1,r_2,...,r_q are the roots of the polynomial. The last equation is equivalent to

$$\epsilon_t = \frac{1}{(1 - k_1 \mathrm{B})(1 - k_2 \mathrm{B}) \ldots (1 - k_q \mathrm{B})} Z_t \tag{2.6.19}$$

where $k_i = \frac{1}{r_i}$.

To obtain the required function in terms of Z_t we need that the RHS of 2.6.19 may be expanded in a convergent power series (see Priestley, 1981, pp. 144-145) with an absolutely summable sequence of parameters, that is

$$\Phi(\mathrm{B}) = \Theta^{-1}(\mathrm{B}) = \sum_{i=0}^{p} \phi_i \mathrm{B}^i < \infty, \qquad \sum_{i=0}^{\infty} \mid \phi_i \mid < \infty. \tag{2.6.20}$$

To satisfy 2.6.20 we need the roots of the polynomial $\Theta(\mathrm{B})$ to be bigger than 1 and equally k_1,k_2,...,k_q smaller than 1. That is the roots have to lie outside the unit circle. This is referred as the invertibility condition. If this is satisfied the Wold representation 2.3.1 of Z_t is identical with the model, 2.6.14 i.e. ϵ_t is the linear innovation of Z_t.

Example 2.6.3 *Consider the MA(1) model 2.6.16. It can be written as*

$$\begin{aligned} \epsilon_t &= Z_t + \theta \epsilon_{t-1} = Z_t + \theta(Z_{t-1} + \theta \epsilon_{t-2}) = \\ &= Z_t + \theta Z_{t-1} + \theta^2 \epsilon_{t-2} = \\ &= Z_t + \theta Z_{t-1} + \theta^2 Z_{t-2} + \ldots + \theta^p Z^{t-p} + \theta^{p+1} \epsilon_{t-p-1} \end{aligned} \tag{2.6.21}$$

which is convergent for $p \to \infty$ *if* $\mid \theta \mid < 1$ *and equally* $\mid r \mid = \frac{1}{\mid \theta \mid} > 1$ *where* r *is the root of* $\Theta(B) = (1 - \theta B)$. □

Example 2.6.4 *Consider the model*

$$Z_t = e_t + 2e_{t-1} \tag{2.6.22}$$

it has

$$\gamma(0) = 5\sigma_e^2; \qquad \gamma(1) = 2\sigma_e^2; \qquad \gamma(k) = 0 \quad , k > 1. \tag{2.6.23}$$

Let us consider now the process 2.3.1

$$Z_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}.$$

Reproducing the covariance structure 2.6.23 for the process 2.3.1 we have

$$\gamma(k) = 0 \ , k > 1 \quad \Rightarrow \quad \psi_j = 0 \ , j > 1$$

hence 2.3.1 becomes

$$Z_t = \epsilon_t + \psi_1 \epsilon_{t-1} \tag{2.6.24}$$

which is characterised by

$$\gamma(0) = (1 + \psi_1^2)\sigma_e^2 \qquad \textit{and} \qquad \gamma(1) = \psi_1 \sigma_e^2. \tag{2.6.25}$$

Considering jointly 2.6.23 and 2.6.25 we have

$$\rho(1) = \frac{\psi_1}{1 + \psi_1^2} = \frac{2}{5} = 0.4$$

which has solutions $\psi = 2$ and $\psi = \frac{1}{2}$. Because of conditions 2.3.3, the Wold representation must have $\psi = \frac{1}{2}$ which, equating $\gamma(1)$ in 2.6.23 and 2.6.25, implies

$$\sigma_\epsilon^2 = \frac{2\sigma_e^2}{\psi_1} = 4\sigma_e^2$$

hence process $\{e_t\}$ in 2.6.22 is different from process $\{\epsilon_t\}$ in 2.3.1. □

2.6.2 Autoregressive models

Approximating the spectrum with polynomials like 2.6.4 leads to

$$Z_t = \frac{1}{(1 + \phi_1 \mathrm{B} + \phi_2 \mathrm{B}^2 + \ldots + \phi_p \mathrm{B}^p)} \epsilon_t$$

giving

$$\Phi(\mathrm{B}) Z_t = (1 - \phi_1 \mathrm{B} - \phi_2 \mathrm{B}^2 - \ldots - \phi_p \mathrm{B}^p) Z_t = \epsilon_t$$

or

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + \epsilon_t \tag{2.6.26}$$

This is called autoregressive model of order p AR(p).

Example 2.6.5 *The first-order autoregressive process AR(1) is given by*

$$Z_t = \phi Z_{t-1} + \epsilon_t. \tag{2.6.27}$$

□

To ensure finite moments of first and second order we need the roots of the polynomial $\Phi(\mathrm{B})$ to lie outside the unit circle. This is called stationarity condition.

Example 2.6.6 *Consider*

$$\begin{aligned} Z_t &= \phi_1 Z_{t-1} + \epsilon_t = \\ &= \phi(\phi Z_{t-2} + \epsilon_{t-1}) + \epsilon_t = \\ &= \epsilon_t + \phi\epsilon_{t-1} + \phi^2 Z_{t-2} = \\ &= \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \ldots + \phi^p\epsilon_{t-p} + \phi^{p+1} Z_{t-p-1}. \end{aligned} \tag{2.6.28}$$

Then mean and variance are

$$\mu_Z = \phi^{p+1} Z_{t-p-1}; \qquad \sigma_Z^2 = \sum_{i=0}^{p} \phi^{2i}\sigma_\epsilon^2 + \phi^{p+1}\sigma_{Z(t-p-1)}^2 \tag{2.6.29}$$

and in order to have them finite for $p \to \infty$ we need the stationarity condition $\mid \phi \mid < 1$. Then Z_t generated by this model is a stationary process with ϵ_t as the linear innovation of Z_t.

□

A MA process is always stationary while an AR process is always invertible.

As examples 2.6.2 and 2.6.4 suggest, a stationary AR model of any order can be represented by a MA model of infinite order and, conversely, an invertible MA model of any order can be represented by an AR model of infinite order. In practice the representations of infinite order can be approximated by representations of finite order.

2.6.3 Mixed models

Approximating the spectrum with a rational polynomial like 2.6.5 we have

$$Z_t = \frac{(1 - \theta_1\mathrm{B} - \theta_2\mathrm{B}^2 - \ldots - \theta_q\mathrm{B}^q)}{(1 - \phi_1\mathrm{B} - \phi_2\mathrm{B}^2 - \ldots - \phi_p\mathrm{B}^p)}\epsilon_t = \frac{\Theta(\mathrm{B})}{\Phi(\mathrm{B})}\epsilon_t \tag{2.6.30}$$

that is

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}. \quad (2.6.31)$$

This kind of model where both AR and MA components are present is called a *mixed models*. In particular 2.6.27 is an autoregressive moving average model of order p and q and is indicated as ARMA(p,q).

Example 2.6.7 *The simplest autoregressive moving average model is ARMA(1,1) and it is given by*

$$Z_t - \phi Z_{t-1} = \epsilon_t - \theta \epsilon_{t-1}. \quad (2.6.32)$$

□

ARMA models have the properties examined for MA and AR models, hence they are respectively invertible and stationary if the roots of the polynomials $\Theta(\mathrm{B})$ and $\Phi(\mathrm{B})$ lie outside the unit circle. As can be guessed from what we have said before an ARMA model can be seen as a general model which includes AR and MA models as special cases and given the number of parameters it has a greater capacity of approximation than AR and MA models.

Integrated models arise when an ARMA model of the form $\Upsilon(\mathrm{B})Z_t = \Theta(\mathrm{B})\epsilon_t$ is locally stationary, that is if its stochastic behaviour is invariant to the addition of constant to its level or more generally to the addition of an arbitrary polynomial trend. In this case we have

$$\Upsilon(\mathrm{B})(Z_t + c) = (1 - \upsilon_1 \mathrm{B} - \upsilon_2 \mathrm{B}^2 - \ldots - \upsilon_{p+d} \mathrm{B}^{p+d})(Z_t + c) = \Upsilon(\mathrm{B})(Z_t)$$

which implies

$$\Upsilon(\mathrm{B})c = (1 - \upsilon_1 \mathrm{B} - \upsilon_2 \mathrm{B}^2 - \ldots - \upsilon_{p+d} \mathrm{B}^{p+d})c = 0 \quad (2.6.33)$$

and because c is a constant, we can write 2.6.33 as

$$\Upsilon(1) = (1 - \upsilon_1 - \upsilon_2 - \ldots - \upsilon_{p+d})c = 0 \leftrightarrow \sum_{i=1}^{p+d} \upsilon_i = 1.$$

This implies that there are one or more unit roots of $\Upsilon(\mathrm{B})$, that is

$$\Upsilon(\mathrm{B}) = (1 - \mathrm{B})^d \Phi(\mathrm{B}) \qquad d \geq 1.$$

We can write this kind of model as

$$\Phi(\mathrm{B})(1 - \mathrm{B})^d Z_t = \Theta(\mathrm{B})\epsilon_t \quad (2.6.34)$$

and it is called an *autoregressive integrated moving average model* of order p,d,q ARIMA(p,d,q) where d is the order of integration. Using the difference operator $\nabla^d = (1 - \mathrm{B})^d$ 2.6.34 can also be written as

$$\Phi(\mathrm{B})\nabla^d = \Theta(\mathrm{B})\epsilon_t. \tag{2.6.35}$$

Example 2.6.8 *An ARIMA(0,1,1) is given by*

$$\nabla Z_t = (1 - \theta B)\epsilon_t \tag{2.6.36}$$

which can also be written as

$$\begin{aligned} \epsilon_t &= (1 - \theta B)^{-1}\nabla Z_t = \\ &= (1 + \theta B + \theta^2 B^2 + \ldots)(1 - B)Z_t = \\ &= (1 - \upsilon_1 B - \upsilon_2 B^2 - \ldots)Z_t \end{aligned} \tag{2.6.37}$$

where $\upsilon_j = (1 - \theta)\theta^{j-1}$, *j=1,2,.... We can rearrange 2.6.37 as*

$$Z_t = \upsilon_1 Z_{t-1} + \upsilon_2 Z_{t-2} + \ldots + \epsilon_t \tag{2.6.38}$$

because

$$\sum_i \upsilon_i = \lim_{j\to\infty}(1 - \theta)\frac{1 - \theta^{j-1}}{1 - \theta} = 1$$

in 2.6.38 Z_t *is expressed an an exponentially weighted average of its past values plus a white noise. This model gives the exponential smoothing predictor.* □

Example 2.6.9 *ARIMA(0,1,0) is represented by the equation*

$$\nabla Z_t = \epsilon_t$$

that is

$$Z_t = Z_{t-1} + \epsilon_t$$

which is a random walk. □

Box and Jenkins (1976) proposed a procedure of three steps, which is now well established, for ARMA modelling:
1) the identification of the specific ARMA model;
2) the estimation of parameters;
3) the checking of the model.

In the next two paragraphs we shall describe the first two steps in more detail while we shall not discuss integrated models.

2.7 Identification of ARMA models

There are two main approaches to the identification of ARIMA models: one is based on the presence of singularities in indicators such as the *sample autocorrelation function* (SACF), the *sample partial autocorrelation function* (SPACF) and the covariance matrix of forecast values; the other one consists of automatic methods (AIC, SCH, HAN, etc...) to compare different models. In this paragraph we shall describe the first approach and in the fifth chapter the second one.

The procedure to identify MA and AR models is well established while several methods were proposed to identify mixed models even if most of them pursue a closely related strategy. For this reason we shall treat separately the identification for ARMA models and that for pure MA and AR models.

2.7.1 MA and AR models

These models are identified by considering the theoretical ACF and PACF and comparing them with the behaviour of the *sample autocorrelation function* (SACF). The sample autocovariance function is given by

$$\hat{\gamma}(k) = \frac{\sum_{t=1}^{n-k}(x_t - \hat{\mu}_x)(x_{t+k} - \hat{\mu}_x)}{n} \tag{2.7.1}$$

where $\{x_t\}$ is the observed time series, n is the number of observations and $\hat{\mu_x}$ the sample mean. The *sample autocorrelation function* (SACF), given by

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}. \tag{2.7.2}$$

The autocovariance function for a MA(q) is given by

$$\gamma(k) = \begin{cases} \sigma^2 \sum_{j=0}^{q-k} \theta_j \theta_{j+k}, & k = 0, 1, 2, \ldots, q \\ 0, & k = q+1, q+2, \ldots \end{cases} \tag{2.7.3}$$

where $\theta_0 = -1$, $\theta_j = 0$, $j > q$. In particular the variance of Z_t will be

$$\gamma(0) = \sigma^2(1 + \theta_1^2 + \ldots + \theta_q^2). \tag{2.7.4}$$

Hence the autocorrelation function will be

$$\rho(k) = \begin{cases} \frac{\sum_{j=0}^{q-k} \theta_j \theta_{j+k}}{\sum_{j=0}^{q} \theta_j^2}, & k = 0, 1, 2, \ldots, q \\ 0, & k = q+1, q+2, \ldots \end{cases} \tag{2.7.5}$$

Thus the characteristic feature of a MA(q) process is that its autocorrelation function becomes zero from lag q onwards. This feature enables us to identify a MA(q) by observing the maximum lag q at which the SACF is significantly different from zero.

As we have seen before a stationary AR model of any order can be represented by a MA of infinite order. Hence any stationary AR model has no such maximum lag of non zero ACF, although there will be a, usually large, lag beyond which the SACF is not significantly different from zero. This implies that the ACF cannot help us in determining the order of an AR model, which can be determined by the PACF. The characteristic feature of an AR process is, in contrast, that the partial autocorrelation $\pi(k)$ becomes zero from lag $p+1$ onwards. Consider

$$\text{Cov}\left(Z_t, Z_{t-k} \mid Z_{t-1}, \ldots, Z_{t-k-1}\right) \qquad Z_t \sim AR(p), k > p \tag{2.7.6}$$

which is equivalent to

$$\text{Cov}\left(Z_t - \phi_1 Z_{t-1} - \ldots - \phi_p Z_{t-p}, Z_{t-k} \mid Z_{t-1}, \ldots, Z_{t-k-1}\right) \tag{2.7.7}$$

because $Z_{t-1} \ldots Z_{t-p}$ are within the set of conditioning variables $Z_{t-1}, \ldots, Z_{t-k-1}$ provided $k > p$. This covariance is

$$\text{Cov}\left(\epsilon_t, Z_{t-k} \mid Z_{t-1}, \ldots, Z_{t-k-1}\right) \tag{2.7.8}$$

but for the AR(p) model ϵ_t is uncorrelated with all past values including $Z_{t-k}, Z_{t-1}, \ldots, Z_{t-k-1}$, so that covariance is zero, and hence so is the correlation $\pi(k)$; and from equation 2.5.36 we know that for the last coefficient in the AR(p) model $\pi(p) = \phi(p) \neq 0$.

2.7.2 ARMA models

There are many methods to identify ARMA models and all eventually rely upon the *extended Yule-Walker equations*. For the ARMA(p,q) models,

$$Z_t - \phi_1 Z_{t-1} - \phi_2 Z_{t-2} - \ldots - \phi_p Z_{t-p} = \epsilon_t - \theta_1 \epsilon_{t-1} - \ldots - \theta_q \epsilon_{t-q} \tag{2.7.9}$$

we extend the argument in 2.5.38 to obtain 2.5.42, but for $k > q$ because only for $k > q$ the covariance between Z_{t-k} and the RHS is equal to zero. Therefore

$$\rho(k) = \phi_1 \rho(k-1) + \ldots + \phi_p \rho(k-p) \qquad \text{, for } k > q. \tag{2.7.10}$$

Note that if $p = 0$ we get the characteristic property of the MA(q) that $\gamma(k) = 0$ for $k > q$. If $q = 0$ we have the standard Yule-Walker equations for the AR(p). Using

sample values r_k of ρ_k these equations may be used to determine estimates of $\phi_1, \ldots, \phi_p$ and the sample ACF of the process $W_t = Z_t - \hat{\phi}_1 Z_{t-1} - \ldots - \hat{\phi}_p Z_{t-p}$ computed. If this is done for an increasing value of p, then when the true value of p is reached W_t will be a MA(q) process recognised by its characteristic property.

2.8 Estimation of ARMA models

The main statistical inference procedures used to estimate models are Bayes and likelihood inference. In practice these two procedures give similar result and both rely on determining the parameters which minimise the sum of squares of the one step ahead prediction errors. The simplest and most reliable assumption to make, when dealing with linear processes, is that the time series are Gaussian. Estimation methods based on this assumption have good properties even if the multivariate normal distribution doesn't describe perfectly the data. In the next subsection we shall explain the maximum likelihood approach as applied to the estimation of AR, MA and ARMA models.

2.8.1 AR models

Consider the AR(1) model

$$z_t = \phi z_{t-1} + \epsilon_t \qquad \text{, for } t = 1, \ldots, n$$

rearranging we can obtain the prediction error

$$\epsilon_t = z_t - \phi z_{t-1} \qquad \text{, for } t = 2, \ldots, n.$$

Hence because of the mutual independence of the errors we can write the joint p.d.f. of the observed time series $\{z_1, z_2, \ldots, z_n\}$ as

$$\mathrm{f}(z_1, z_2, \ldots, z_n) = \mathrm{f}(z_1)\mathrm{f}(\epsilon_2)\mathrm{f}(\epsilon_3)\ldots\mathrm{f}(\epsilon_n) \propto \mathrm{f}(z_1)\sigma_\epsilon^{-(n-1)}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left(\sum_{t=2}^{n}\epsilon_t^2\right)\right\}. \quad (2.8.1)$$

If we consider z_1 as a fixed quantity which considered alone does not contribute to the information needed to estimate ϕ, that is $\mathrm{f}(z_1) = k$, then the likelihood is

$$\mathrm{L}(\phi) = k\sigma_\epsilon^{-(n-1)}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left(\sum_{t=2}^{n}\epsilon_t^2\right)\right\} \quad (2.8.2)$$

and the log-likelihood

$$\ell(\phi) = \log k - (n-1)\log\sigma_\epsilon - \frac{1}{2\sigma_\epsilon^2}\left(\sum_{t=2}^{n}\epsilon_t^2\right) \quad (2.8.3)$$

which is maximised when we minimise the sum of the squared residuals

$$S = \sum_{t=2}^{n} \epsilon_t^2 \tag{2.8.4}$$

which is the standard least square regression. Omitting the constant term it is of interest to compare then least-squares estimate with Yule-Walker estimate. The parameter ϕ estimated by least squares regression is given by

$$\begin{aligned} \frac{\partial S}{\partial \phi} &= \frac{\partial \sum_{t=1}^{n} \epsilon_t^2}{\partial \phi} \\ &= \frac{\partial \sum_{t=1}^{n} (z_t - \phi z_{t-1})^2}{\partial \phi} \\ &= \sum_{t=1}^{n} (\phi z_{t-1}^2 - z_t z_{t-1}) = 0 \end{aligned} \tag{2.8.5}$$

which implies

$$\hat{\phi} = \frac{\sum_{t=1}^{n} z_t z_{t-1}}{\sum_{t=1}^{n} z_{t-1}^2} \tag{2.8.6}$$

which can be ≥ 1, while the Yule-Walker estimator is given by

$$\begin{aligned} \hat{\phi} &= \hat{\rho}(1) \\ &= \frac{\sum_{t=2}^{n} z_t z_{t-1}}{\sum_{t=1}^{n} z_t^2} < 1 \end{aligned} \tag{2.8.7}$$

which hence always gives stationary models.

In order to get an exact maximum likelihood estimation we need to take into account the information z_1 and not treat it as a constant. In this case the likelihood is

$$\mathrm{L}(\phi) = \mathrm{f}(z_1)\sigma_\epsilon^{-(n-1)} \exp\left\{ -\frac{1}{2\sigma_\epsilon^2} \left(\sum_{t=2}^{n} \epsilon_t^2 \right) \right\}. \tag{2.8.8}$$

Considering that the variance of an AR(1) process is

$$\begin{aligned} \mathrm{Var}\,(z_t) = \sigma_z^2 &= \mathrm{Var}\,(\phi z_{t-1} + \epsilon_t) \\ &= \mathrm{Var}\,(\epsilon + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + + \ldots) \\ &= \sigma_\epsilon^2 + \phi^2\sigma_\epsilon^2 + \phi^4\sigma_\epsilon^2 + \ldots \\ &= \sigma_\epsilon^2 \left(\frac{1}{1-\phi^2} \right) \end{aligned} \tag{2.8.9}$$

the p.d.f. of z_1 is

$$\begin{aligned} \mathrm{f}(z_1) &= \frac{1}{2\pi\sigma_z} \exp\left\{ -\frac{1}{2\sigma_z^2} z_1^2 \right\} \\ &\propto \frac{1}{\sigma_z} \exp\left\{ -\frac{1}{2\sigma_z^2} z_1^2 \right\} \\ &\propto \frac{\sqrt{(1-\phi^2)}}{\sigma_\epsilon} \exp\left\{ -\frac{(1-\phi^2)}{2\sigma_\epsilon^2} z_1^2 \right\}. \end{aligned} \tag{2.8.10}$$

Hence 2.8.8 becomes

$$\mathrm{L}(\phi) = \sqrt{(1-\phi^2)}\sigma_\epsilon^{-n}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left[(1-\phi^2)z_1^2 + \sum_{t=2}^{n}\epsilon_t^2\right]\right\} \tag{2.8.11}$$

and

$$\ell(\phi) = -n\log(\sigma_\epsilon) - \frac{1}{2}\log(1-\phi^2) - \frac{1}{2\sigma_\epsilon^2}\left[(1-\phi^2)z_1^2 + \sum_{t=2}^{n}\epsilon_t^2\right] \tag{2.8.12}$$

which is maximised when we minimise

$$\log(1-\phi^2) + \frac{1}{\sigma_\epsilon^2}\left[(1-\phi^2)z_1^2 + \sum_{t=2}^{n}\epsilon_t^2\right] \tag{2.8.13}$$

which requires a non linear least squares procedure.

Nevertheless a nonlinear least squares convergences quickly to give the estimate of ϕ. The reasoning can be easily extended to AR models of higher degree.

2.8.2 MA models

Similarly for a MA(1) model we can calculate the prediction errors ϵ_t from the data z_t, for $t = 2, 3, \ldots, n$. The calculation is recursive, that is

$$\epsilon_t = z_t + \theta\epsilon_{t-1} \qquad , t = 1, 2, 3, \ldots, n \tag{2.8.14}$$

and the consequent joint p.d.f. of the data is

$$\mathrm{f}(\epsilon_0, \epsilon_1, \ldots, \epsilon_n) = \mathrm{f}(\epsilon_0, z_1, \ldots, z_n) = \mathrm{f}(\epsilon_0)\mathrm{f}(\epsilon_1)\ldots\mathrm{f}(\epsilon_n) \propto \sigma_\epsilon^{-(n+1)}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}S\right\} \tag{2.8.15}$$

where $S = \sum_{t=0}^{n+1}\epsilon_t^2$ and minimising it we should obtain the maximum likelihood estimation. Nevertheless we have the problem that ϵ_0 is unknown. A simple strategy is to set it to zero, this corresponds to set $\epsilon_1 = z_1$. This is a common practice but if the value of θ was close to one, this strategy could distort the predictions for the residuals for several time points. Hence as soon as we have enough data it would make sense to use a better *back forecast* predictor for z_0. Letting $\epsilon_0 = 0$ in the MA(1) model

$$\begin{aligned}\hat{z}_t &= -\theta\epsilon_{t-1} \\ &= -\theta(1-\theta\mathrm{B})^{-1}\hat{z}_{t-1} \\ &= -\theta\hat{z}_{t-1} - \theta^2\hat{z}_{t-2} - \theta^3\hat{z}_{t-3}\ldots\end{aligned} \tag{2.8.16}$$

also

$$\epsilon_0 = (1-\theta\mathrm{B})^{-1}z_0 = z_0 + \theta z_{-1} + \theta^2 z_{-2} + \ldots \tag{2.8.17}$$

Nevertheless $z_{-1}, z_{-2}, \ldots$ are unknown and if we let them, consistently with what we did for z_1, be equal to $\epsilon_{-1}, \epsilon_{-2}, \ldots = 0$ we have

$$\hat{\epsilon}_0 = z_0$$

and from 2.8.16 by time reversibility and generalising

$$\hat{z}_t = \hat{\epsilon}_t = -\theta z_{t+1} - \theta^2 z_{t+2} - \theta^3 z_{t+3} \ldots \tag{2.8.18}$$

Because of the *Gauss-Markov theorem* (see Azzalini, 1996, pp. 179–180) the same estimate can be obtained as a least squares estimate minimising S where

$$\begin{aligned} S &= \sum_{t=1}^{n+1} \epsilon_t^2 \\ &= \epsilon_0^2 + \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 + \ldots \\ &= \epsilon_0^2 + (z_1 + \theta\epsilon_0)^2 + [z_2 + \theta(z_1 + \theta\epsilon_0)]^2 + \{z_3 + \theta[z_2 + \theta(z_1 + \theta\epsilon_0)]\}^2 + \ldots \\ &= \epsilon_0^2 + (z_1 + \theta\epsilon_0)^2 + (z_2 + \theta z_1 + \theta^2\epsilon_0)^2 + (z_3 + \theta z_2 + \theta^2 z_1 + \theta^3\epsilon_0)^2 + \ldots \end{aligned}$$

hence

$$\begin{aligned} \frac{\partial S}{\partial \epsilon_0} &= 2 \quad \Big[\epsilon_0 + (z_1 + \theta\epsilon_0)\theta + (z_2 + \theta z_1 + \theta^2\epsilon_0)\theta^2 + (z_3 + \theta z_2 + \theta^2 z_1 + \theta^3\epsilon_0)\theta^3 + \ldots\Big] \\ &= 2 \quad \Big[(1 + \theta^2 + \theta^4 + \ldots + \theta^{2n})\epsilon_0 \\ &\quad + \quad (\theta + \theta^3 + \theta^5 + \ldots + \theta^{2(n-1)+1})z_1 \\ &\quad + \quad (\theta^2 + \theta^4 + \theta^6 + \ldots + \theta^{2(n-1)})z_2 \\ &\quad + \quad (\theta^3 + \theta^5 + \theta^7 \ldots + \theta^{2(n-2)+1})z_3 \\ &\quad + \quad .....\Big] \end{aligned}$$

then equating $\frac{\partial S}{\partial \epsilon_0} = 0$ and letting $1 + \theta^2 + \theta^4 + \ldots + \theta^{2n} = K$, we have

$$\begin{aligned} K\epsilon_0 = \; - \quad & \Big\{\theta\frac{1 - \theta^{2(n-1)}}{1 - \theta^2}z_1 \\ + \quad & \theta^2\frac{1 - \theta^{2(n-2)}}{1 - \theta^2}z_2 \\ + \quad & \theta^3\frac{1 - \theta^{2(n-3)}}{1 - \theta^2}z_3 \\ + \quad & .....\Big\} \end{aligned}$$

then substituting for $z_1, z_2, z_3, \ldots$ and letting $\epsilon_0 = 0$

$$K\epsilon_0 = \; - \quad \Big\{\theta\frac{1 - \theta^{2(n-1)}}{1 - \theta^2}\epsilon_1$$

$$
\begin{aligned}
&+ \quad \theta^2 \frac{1-\theta^{2(n-2)}}{1-\theta^2}(\epsilon_2 - \theta\epsilon_1) \\
&+ \quad \theta^3 \frac{1-\theta^{2(n-3)}}{1-\theta^2}(\epsilon_3 - \theta\epsilon_2) \\
&+ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\Bigg\}
\end{aligned}
$$

and rearranging we obtain

$$
\begin{aligned}
K\epsilon_0 = \quad - \quad &\left\{\left[\theta\frac{1-\theta^{2(n-1)}}{1-\theta^2} - \theta^3\frac{1-\theta^{2(n-2)}}{1-\theta^2}\right]\epsilon_1\right. \\
&+ \quad \left[\theta^2\frac{1-\theta^{2(n-2)}}{1-\theta^2} - \theta^4\frac{1-\theta^{2(n-3)}}{1-\theta^2}\right]\epsilon_2 \\
&+ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\Bigg\}
\end{aligned}
$$

and finally solving

$$\hat{\epsilon}_0 = -\left(\frac{\theta\epsilon_1 + \theta^2\epsilon_2 + \ldots + \theta^n\epsilon_n}{K}\right). \tag{2.8.19}$$

The decomposition of sum of squares for this regression gives

$$S = K(\epsilon_0 - \hat{\epsilon}_0) + \hat{S} \tag{2.8.20}$$

where $\hat{S}$ is the minimum value of S got by using $\hat{\epsilon}_0$ to start up the calculations for $\epsilon_1, \epsilon_2, \ldots$. In this way we obtain the conditional mean, ϵ_0, and the conditional variance, $\frac{\sigma_\epsilon^2}{K}$, of ϵ_0 given $z_1, z_2, \ldots, z_n$. Then the p.d.f. 2.8.15 can be written as

$$\mathrm{f}(\epsilon_0, z_1, z_2, \ldots, z_n) = \mathrm{f}(\epsilon_0 \mid z_1, z_2, \ldots, z_n)\mathrm{f}(z_1, z_2, \ldots, z_n) \tag{2.8.21}$$

where

$$\mathrm{f}(\epsilon_0, z_1, z_2, \ldots, z_n) \propto \frac{K^{\frac{1}{2}}}{\sigma_\epsilon}\exp\left\{-\frac{K}{2\sigma_\epsilon^2}(\epsilon_0 - \hat{\epsilon}_0)^2\right\} \tag{2.8.22}$$

hence from 2.8.15, 2.8.21 and 2.8.22 we have

$$\mathrm{f}(z_1, z_2, \ldots, z_n) =\propto K^{-\frac{1}{2}}\sigma_\epsilon^{-n}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\hat{S}\right\}. \tag{2.8.23}$$

Because the terms in S do not depend linearly upon θ, iterative non-linear least squares methods must be used to obtain the maximum likelihood estimates. The reasoning can be extended to a general MA(q) model by back forecast $\epsilon_0, \epsilon_1, \ldots, \epsilon_{q-1}$.

2.8.3 ARMA models

With a similar procedure used for both the AR and the MA model we now derive the maximum likelihood estimate for ARMA model, considering the simple case of an

ARMA(1,1), that is

$$z_t - \phi z_{t-1} = \epsilon_t - \theta\epsilon_{t-1} \tag{2.8.24}$$

and expressing z_t as a function of the innovations

$$z_t = \frac{(1-\theta \mathrm{B})}{(1-\phi \mathrm{B})}\epsilon_t = \epsilon_t + (\phi-\theta)\epsilon_{t-1} + \phi(\phi-\theta)\epsilon_{t-2} + \phi^2(\phi-\theta)\epsilon_{t-3} + \ldots \tag{2.8.25}$$

To obtain a recursive computation of the errors we need to rearrange 2.8.24 as

$$\epsilon_t = z_t - \phi z_{t-1} + \theta\epsilon_{t-1} \tag{2.8.26}$$

and again we have to set a starting value for $\{\epsilon_0\}$, letting it equal to zero, the conditional expectation of z_t will be

$$\begin{aligned} \mathrm{E}(z_t \mid \epsilon_t) = z_t - \epsilon_t &= \phi z_{t-1} - \theta\epsilon_{t-1} \\ &= (\phi-\theta)\epsilon_{t-1} + \phi(\phi-\theta)\epsilon_{t-2} + \phi^2(\phi-\theta)\epsilon_{t-3} + \ldots \end{aligned} \tag{2.8.27}$$

and the conditional variance

$$\begin{aligned} \mathrm{Var}\,(z_t \mid \epsilon_t) &= \left[(\phi-\theta)^2 + \phi^2(\phi-\theta)^2 + \phi^4(\phi-\theta)^2 + \ldots\right]\sigma_\epsilon^2 \\ &= \frac{(\phi-\theta)^2}{(1-\phi^2)}\sigma_\epsilon^2 \end{aligned} \tag{2.8.28}$$

letting $\delta^2 = \frac{(1-\phi^2)}{(\phi-\theta)^2}$, the joint p.d.f. of the data will be:

$$\mathrm{f}(\epsilon_1, z_1, z_2, \ldots, z_n) = \mathrm{f}(z_1 \mid \epsilon_1)\mathrm{f}(\epsilon_1)\mathrm{f}(\epsilon_2), \ldots, \mathrm{f}(\epsilon_n) \propto \delta\sigma_\epsilon^{-n}\exp\left\{-\frac{1}{2\sigma_\epsilon^2}S\right\} \tag{2.8.29}$$

and minimising S with respect to ϵ_1, now we have

$$\hat{\epsilon}_1 = \frac{\delta^2(z_1-\epsilon_1) - \theta\epsilon_2 - \theta^2\epsilon_3 - \ldots - \theta^{n-1}\epsilon_n}{K} \tag{2.8.30}$$

where $\epsilon_2, \epsilon_3, \ldots$ are the residuals computed starting up with $\epsilon_1 = 0$ and

$$K = \delta^2 + 1 + \theta^2 + \theta^4 + \ldots + \theta^{2(n-1)}$$

and using a similar argument as for the MA(1) we therefore identify the likelihood expression.

Chapter 3

Multivariate Time Series Analysis

In this chapter we extend to the multivariate case some of the concepts examined in chapter 2.

3.1 From multivariate stochastic processes to vector ARMA models

The process $\{\boldsymbol{Y}_t\}$ defined by the vector

$$\{\boldsymbol{Y}_t\} = \begin{bmatrix} \{Y_{1,t}\} \\ \{Y_{2,t}\} \\ \vdots \\ \{Y_{k,t}\} \end{bmatrix} \tag{3.1.1}$$

where $\{Y_{1,t}\}, \{Y_{2,t}\}, \ldots, \{Y_{k,t}\}$ are stochastic processes as seen in chapter 2, is called multivariate stochastic process. It is strictly stationary if the probability distributions of the sets of vectors $(\{\boldsymbol{Y}_1\}, \{\boldsymbol{Y}_2\}, \ldots, \{\boldsymbol{Y}_k\})^T$ and $(\{\boldsymbol{Y}_{1+l}\}, \{\boldsymbol{Y}_{2+l}\}, \ldots, \{\boldsymbol{Y}_{k+l}\})^T$ are the same for any $k = \pm 1, \pm 2, \ldots$ and $l = 0, \pm 1, \pm 2, \ldots$. For weak stationarity, if first and second moments exist, we need

$$\mathrm{E}(\boldsymbol{Y}_t) = \boldsymbol{\mu}$$

for any t, where $\boldsymbol{\mu}^T = (\mu_1, \mu_2, \ldots, \mu_k)$ and the covariance, $\boldsymbol{\Gamma}(l)$, between $\boldsymbol{Y}_t$ and $\boldsymbol{Y}_{t+l}$ depends just on the lag l, that is

$$\boldsymbol{\Gamma}(l) = \mathrm{E}\left[(\boldsymbol{Y}_t - \boldsymbol{\mu})(\boldsymbol{Y}_{t+l} - \boldsymbol{\mu})^T\right] = \begin{bmatrix} \gamma_{11}(l) & \gamma_{12}(l) & \ldots & \gamma_{1k}(l) \\ \gamma_{21}(l) & \gamma_{22}(l) & \ldots & \gamma_{2k}(l) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k1}(l) & \gamma_{k2}(l) & \ldots & \gamma_{kk}(l) \end{bmatrix} \tag{3.1.2}$$

where $\gamma_{ij}(l) = \text{Cov}\,(Y_{i,t}, Y_{j,t+l})$, which has the property

$$\boldsymbol{\Gamma}(-l) = \text{E}[(\boldsymbol{Y}_{t+l} - \boldsymbol{\mu})(\boldsymbol{Y}_t - \boldsymbol{\mu})^T] = \boldsymbol{\Gamma}(l)^T \tag{3.1.3}$$

since $\gamma_{i,j}(l) = \gamma_{i,j}(-l)$ and which is non-negative definite.

Suppose we have a stationary process $\{\boldsymbol{X}_t\}$ corrected for any deterministic component of the Wold decomposition. Consider the regression on all past values

$$\boldsymbol{X}_t = \boldsymbol{\Pi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\Pi}_2 \boldsymbol{X}_{t-2} + \ldots + \boldsymbol{E}_t. \tag{3.1.4}$$

This defines

$$\boldsymbol{E}_t = \boldsymbol{X}_t - \boldsymbol{\Pi}_1 \boldsymbol{X}_{t-1} - \boldsymbol{\Pi}_2 \boldsymbol{X}_{t-2} - \ldots = \boldsymbol{\Pi}(\text{B})\boldsymbol{X}_t \tag{3.1.5}$$

as the vector of *linear innovations* in $\boldsymbol{X}_t$ from the whole past, hence $\boldsymbol{E}_t$ is a stationary white noise process such that $\boldsymbol{E}_t \perp \boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots$ and $\boldsymbol{E}_t \perp \boldsymbol{E}_{t-1}, \boldsymbol{E}_{t-2}, \ldots$ It is possible to represent $\boldsymbol{X}_t$ by

$$\boldsymbol{X}_t = \boldsymbol{E}_t + \boldsymbol{\Psi}_1 \boldsymbol{E}_{t-1} + \boldsymbol{\Psi}_2 \boldsymbol{E}_{t-2} + \ldots = \boldsymbol{\Psi}(\text{B})\boldsymbol{E}_t \tag{3.1.6}$$

which is the inverse of 3.1.4 in the sense that $\boldsymbol{\Psi}(\text{B})\boldsymbol{\Pi}(\text{B}) = \boldsymbol{I}$. This is the stochastic component of the Wold decomposition for multivariate stationary processes.

Besides the coefficients $\boldsymbol{\Psi}(\text{B})$ and $\boldsymbol{\Pi}(\text{B})$, an important parameter of $\boldsymbol{X}_t$ is the covariance matrix of the innovations

$$\text{Var}\,(\boldsymbol{E}_t) = \boldsymbol{V}$$

which is uniquely defined and has the only constraint of being positive definite. In particular, the contemporaneous elements of $\boldsymbol{E}_t$ will generally be correlated. The variance of $\boldsymbol{X}_t$ may be expressed as $\text{Var}\,(\boldsymbol{X}_t) = \sum_j \boldsymbol{\Psi}_j \boldsymbol{V} \boldsymbol{\Psi}_j^T$ which demonstrates that $\boldsymbol{\Psi}_j$ is squared summable (see Hannan, 1970, 137–157). The stronger assumption that $\boldsymbol{\Psi}_j$ is absolutely summable is also commonly made.

If the series in 3.1.4 is finite it is a vector AR (VAR). In practice a finite vector AR will adequately approximate any process if it is of sufficiently high order.

3.2 Multivariate AR model

Vector autoregressions (VAR) in the time series literature (e.g. Lütkepohl, 1993) relate present observations of each variable to past values of all the variables up to some

maximum lag p, the (overall) order of the VAR and we shall call this the canonical VAR. It does not include any contemporaneous relationship between the elements of $\boldsymbol{X}_t$. Models which also include such relationships, besides lagged variables, are useful in time series modelling and are known as structural VAR models (see Hendry, 1995, 314–315).

3.2.1 Canonical VAR model

Similarly to the univariate case, a vector autoregressive model of order p is represented by

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{X}_{t-2} + \ldots + \boldsymbol{\Phi}_p \boldsymbol{X}_{t-p} + \boldsymbol{E}_t \tag{3.2.1}$$

where $\mathrm{Var}\,(\boldsymbol{E}_t) = \boldsymbol{V}$.

Example 3.2.1 *Consider the VAR(2) model*

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{X}_{t-2} + \boldsymbol{E}_t$$

that is

$$\begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix} = \begin{bmatrix} \phi_{1,1,1} & \phi_{1,2,1} \\ \phi_{2,1,1} & \phi_{2,2,1} \end{bmatrix} \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \end{bmatrix} + \begin{bmatrix} \phi_{1,1,2} & \phi_{1,2,2} \\ \phi_{2,1,2} & \phi_{2,2,2} \end{bmatrix} \begin{bmatrix} X_{1,t-2} \\ X_{2,t-2} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}$$

which represents the equation system

$$\begin{cases} X_{1,t} = & \phi_{1,1,1}X_{1,t-1} & + & \phi_{1,2,1}X_{2,t-1} & + & \phi_{1,1,2}X_{1,t-2} & + & \phi_{1,2,2}X_{2,t-2} + \epsilon_{1,t} \\ X_{2,t} = & \phi_{2,1,1}X_{1,t-1} & + & \phi_{2,2,1}X_{2,t-1} & + & \phi_{2,1,2}X_{1,t-2} & + & \phi_{2,2,2}X_{2,t-2} + \epsilon_{2,t}. \end{cases}$$

Parameter matrices can be sparse, hence the equation system

$$\begin{cases} x_t = & 1.1x_{t-1} & - & 0.2x_{t-2} & + & e_t \\ y_t = & 0.2x_{t-1} & + & 0.7y_{t-1} & + & f_t \end{cases} \tag{3.2.2}$$

can be represented by the VAR(2)

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 1.1 & 0 \\ 0.2 & 0.7 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} -0.2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{t-2} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} e_t \\ f_t \end{bmatrix}.$$

□

Equation 3.2.1 can also be written as

$$(\boldsymbol{I} - \boldsymbol{\Phi}_1 \mathrm{B} - \boldsymbol{\Phi}_2 \mathrm{B}^2 - \ldots - \boldsymbol{\Phi}_p \mathrm{B}^p)\boldsymbol{X}_t = \boldsymbol{E}_t \tag{3.2.3}$$

that is

$$\boldsymbol{\Phi}(\mathrm{B})\boldsymbol{X}_t = \boldsymbol{E}_t \tag{3.2.4}$$

and rearranging

$$\boldsymbol{X}_t = \boldsymbol{\Phi}(\mathrm{B})^{-1}\boldsymbol{E}_t. \tag{3.2.5}$$

From 3.1.6 and 3.2.5 we have

$$\boldsymbol{\Phi}(\mathrm{B})^{-1} = \boldsymbol{\Psi}(\mathrm{B}) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \mathrm{B}^j. \tag{3.2.6}$$

The requirement that $\boldsymbol{\Phi}(\mathrm{B})^{-1}$ converges for $\mid \mathrm{B} \mid \leq 1$ ensures that 3.2.6 converges and 3.2.3 represents a stationary process. The stationarity condition required is that $\det(\boldsymbol{\Phi}(\mathrm{B}))$ has no zeros for $\mid \mathrm{B} \mid \leq 1$.

3.2.2 Structural VAR models

Structural VAR models also allow contemporaneous relationships between the elements of $\boldsymbol{X}_t$ besides lagged relationships. They may therefore be represented as

$$\boldsymbol{\Phi}_0^* \boldsymbol{X}_t = \boldsymbol{\Phi}_1^* \boldsymbol{X}_{t-1} + \ldots + \boldsymbol{\Phi}_p^* \boldsymbol{X}_{t-p} + \boldsymbol{E}_t^*. \tag{3.2.7}$$

The structural VAR where Φ^* has some non-zero off diagonal elements follows as an extension of the simple structural model.

$$\boldsymbol{\Phi}_0^* \boldsymbol{X}_t = \boldsymbol{E}_t^* \tag{3.2.8}$$

by introducing lagged variables.

In contrast with the canonical VAR which has an empirical statistical background, the structural VAR arises from models in which the relationships are based upon econometric modelling purposes and hypotheses. They therefore have associated with them a different set of constraints and conditions.

Typically, the elements of $\boldsymbol{E}_t^*$ are taken to be uncorrelated between themselves and uncorrelated with specified elements of $\boldsymbol{X}_t$. There also may be other constraints that some of the elements of $\boldsymbol{\Phi}_0^*$ are zero. Such constraints are required to ensure that the model is uniquely identified.

The canonical VAR may be viewed as a particular structural VAR with the constraint $\boldsymbol{\Phi}_0^* = \boldsymbol{I}$ and no constraints imposed on $\boldsymbol{E}_t^*$.

Lemma 3.2.2 *Any structural VAR can be transformed to a unique canonical VAR by*

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{\Phi}_0^{*-1}\boldsymbol{\Phi}_1^* \boldsymbol{X}_{t-1} + \ldots + \boldsymbol{\Phi}_0^{*-1}\boldsymbol{\Phi}_p^* \boldsymbol{X}_{t-p} + \boldsymbol{\Phi}_0^{*-1}\boldsymbol{E}_t^* \\ &= \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \ldots + \boldsymbol{\Phi}_p \boldsymbol{X}_{t-p} + \boldsymbol{E}_t \end{aligned} \tag{3.2.9}$$

called the final form. □

Example 3.2.3 *Consider the bivariate process*

$$\begin{cases} x_t = 0.9x_{t-1} + \epsilon_{x_t} \\ y_t = 0.6y_{t-1} + 0.4x_t + \epsilon_{y_t} \end{cases} \tag{3.2.10}$$

where ϵ_{x_t} and ϵ_{y_t} are white noises independent among themselves, distributed according to a N(0,1). It can be represented by a structural VAR

$$\begin{pmatrix} 1 & 0 \\ -0.4 & 1 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.6 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{x_t} \\ \epsilon_{y_t} \end{pmatrix}. \tag{3.2.11}$$

Because of independence the residual covariance matrix is

$$V_\epsilon = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We can represent 3.2.11 as a canonical VAR

$$\begin{aligned} \begin{pmatrix} x_t \\ y_t \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ -0.4 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0.9 & 0 \\ 0 & 0.6 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ -0.4 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \epsilon_{x_t} \\ \epsilon_{y_t} \end{pmatrix} = \\ &= \begin{pmatrix} 0.9 & 0 \\ 0.36 & 0.6 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0.4 & 1 \end{pmatrix} \begin{pmatrix} \epsilon_{x_t} \\ \epsilon_{y_t} \end{pmatrix} \end{aligned} \tag{3.2.12}$$

where the covariance matrix of the errors is given by

$$V_\nu = \begin{pmatrix} 1 & 0 \\ 0.4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0.4 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1.16 \end{pmatrix}. \tag{3.2.13}$$

Systems 3.2.11 and 3.2.12 are stochastically equivalent and the major value of the canonical VAR(p) model is its unique parametrisation. It is also encompassing of all structural VAR(p) models so that for practical applications in prediction it is simple to use without concern for a structural interpretation. □

It is relevant to observe that the equations of a canonical VAR, like 3.2.2, may be separately fitted by regression. However, because the errors e_t, f_t in the equations are in general correlated, i.e. $\mathrm{Var}\,(\boldsymbol{E}_t) = \boldsymbol{V}$, this estimation is not necessarily fully efficient. The *regressions* are then said to be *seemingly unrelated* (SUR).

In many situations a structural VAR(p) model can be expected to be more parsimonious in parametrisation than the corresponding canonical VAR(p) besides having the advantage of interpretability. Unfortunately there are many structural models which result in the same final form canonical model. To identify the structural form uniquely it is therefore necessary to have prior information.

Lemma 3.2.4 *For any given canonical VAR there correspond, in general, many equivalent structural VAR models.*
From equation 3.2.1, let $\boldsymbol{V} = \boldsymbol{TDT}^T$*; many such factorisations exist. Let* $\boldsymbol{\Phi}_0^* = \boldsymbol{T}^{-1}$ *and multiply 3.2.1 by* $\boldsymbol{\Phi}^*$*. This gives the structural VAR (3.2.7) with* $\boldsymbol{\Phi}_j^* = \boldsymbol{\Phi}^*\boldsymbol{\Phi}_j$ *and* $\boldsymbol{E}_t^* = \boldsymbol{\Phi}_0^*\boldsymbol{E}_t$*. By construction* $\mathrm{Var}(\boldsymbol{E}_t) = \boldsymbol{D}$*, so the components of* $\mathrm{Var}(\boldsymbol{E}_t)$ *are uncorrelated.* □

Example 3.2.5 *Matrix 3.2.13 could be given by*

$$V_\nu = \begin{pmatrix} 1 & \frac{0.4}{1.16} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{1.16} & 0 \\ 0 & 1.16 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{0.4}{1.16} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1.16 \end{pmatrix}. \tag{3.2.14}$$

Considering that

$$\begin{pmatrix} 1 & \frac{0.4}{1.16} \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -\frac{0.4}{1.16} \\ 0 & 1 \end{pmatrix}$$

we can rewrite 3.2.12 as

$$\begin{aligned} \begin{pmatrix} x_t \\ y_t \end{pmatrix} \quad &= \quad \begin{pmatrix} 1 & -\frac{0.4}{1.16} \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0.78 & 0.21 \\ 0.36 & 0.6 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} \\ &+ \quad \begin{pmatrix} 1 & -\frac{0.4}{1.16} \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \omega_{x_t} \\ \omega_{y_t} \end{pmatrix} \end{aligned} \tag{3.2.15}$$

where ω_{x_t} *and* ω_{y_t} *are correlated errors with covariance matrix*

$$V_\omega = \begin{pmatrix} \frac{1}{1.16} & 0 \\ 0 & 1.16 \end{pmatrix}.$$

Thus we obtain the structural form

$$\begin{pmatrix} 1 & 3.4 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} 0.78 & 0.21 \\ 0.36 & 0.6 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \omega_{x_t} \\ \omega_{y_t} \end{pmatrix} \tag{3.2.16}$$

which is quite different from 3.2.11. □

Within the class of structural VAR models we can distinguish two main class of submodels: *causal models* and *simultaneous equation models.*

Causal models

These are alternatively known as recursive models.
We have a causal model when it is possible to order system equations in a way for which the contemporaneous coefficient matrix Φ_0^* is triangular.

Example 3.2.6 *Consider the model*

$$\begin{cases} x_t &= \alpha_1 x_{t-1} + \alpha_2 y_t + \alpha_3 z_t + u_t \\ y_t &= \beta_1 y_{t-1} + \beta_2 x_{t-1} + v_t \\ z_t &= \gamma_1 z_{t-1} + \gamma_2 y_t + \gamma_3 x_{t-1} + e_t \end{cases} \tag{3.2.17}$$

with u_t, v_t, e_t uncorrelated. We can order these equations as

$$\begin{cases} y_t &= \beta_1 y_{t-1} + \beta_2 x_{t-1} + v_t \\ z_t &= \gamma_1 z_{t-1} + \gamma_2 y_t + \gamma_3 x_{t-1} + e_t \\ x_t &= \alpha_1 x_{t-1} + \alpha_2 y_t + \alpha_3 z_t + u_t \end{cases} \tag{3.2.18}$$

which can be represented by the matrix form

$$\boldsymbol{\Phi}_0^* \boldsymbol{X}_t = \boldsymbol{\Phi}_1^* \boldsymbol{X}_{t-1} + \boldsymbol{E}_t$$

and more explicitly

$$\begin{bmatrix} 1 & 0 & 0 \\ -\gamma_2 & 1 & 0 \\ \alpha_2 & \alpha_3 & 1 \end{bmatrix} \begin{bmatrix} y_t \\ z_t \\ x_t \end{bmatrix} = \begin{bmatrix} \beta_1 & 0 & \beta_2 \\ 0 & \gamma_1 & \gamma_3 \\ 0 & 0 & \alpha_1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} v_t \\ e_t \\ u_t \end{bmatrix}. \tag{3.2.19}$$

□

The fact that $\boldsymbol{\Phi}_0^*$ is lower triangular implies that every variable is affected only by the preceding contemporaneous variables and so it is possible to determine a causal path which, as we shall see later, in graphical modelling is indicated as a *directed acyclic graph* (DAG). This is the reason why this kind of model is called causal. Econometricians refer to it as a recursive model emphasising the fact that, after the reordering, the first endogenous variable can be estimated straightforwardly from past observations of all the endogenous variables, then the second endogenous variable can be estimated using past observations and the contemporaneous value of the first endogenous variable and so on. It must be noted also that because of the independence of the residuals, their covariance matrix, $\boldsymbol{V}^*$, is diagonal.

Given the ordering there is a unique correspondence between a structural causal model and its correspondent canonical form. In fact

$$\boldsymbol{\Phi}_0^{*-1} \boldsymbol{E}_t^* = \boldsymbol{E}_t \tag{3.2.20}$$

and

$$\boldsymbol{\Phi}_0^{*-1} \boldsymbol{V}^* \boldsymbol{\Phi}_0^{*-T} = \boldsymbol{V} \tag{3.2.21}$$

where $\boldsymbol{V} = \mathrm{Var}\,(\boldsymbol{E}_t)$ in the canonical form can be uniquely express as $\boldsymbol{TDT}^T$ by the *Choleski factorisation*, where $\boldsymbol{T}$ is a lower triangular matrix with unit diagonal elements

$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ t_{2,1} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ t_{n,1} & t_{n,2} & \dots & 1 \end{bmatrix} \tag{3.2.22}$$

and $\boldsymbol{D}$ is the diagonal matrix

$$\boldsymbol{D} = \begin{bmatrix} d_{1,1} & 0 & \dots & 0 \\ 0 & d_{2,1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_{n,n} \end{bmatrix}. \tag{3.2.23}$$

Then we can identify

$$\boldsymbol{\Phi}_0^* = \boldsymbol{T}^{-1} \quad \text{and} \quad \boldsymbol{V}^* = \boldsymbol{D}. \tag{3.2.24}$$

However each ordering of the elements of $\boldsymbol{X}_t$ corresponds to a different structural VAR.

The correlation between $\boldsymbol{X}_t$ and $\boldsymbol{E}_t^*$ satisfies

$$\boldsymbol{\Phi}_0^* \mathrm{Cov}\,(\boldsymbol{X}_t, \boldsymbol{E}_t^*) = \mathrm{Var}\,(\boldsymbol{E}_t^*) = \boldsymbol{V}^* \tag{3.2.25}$$

so

$$\mathrm{Cov}\,(\boldsymbol{X}_t, \boldsymbol{E}_t^*) = \boldsymbol{\Phi}_0^{*-1} \boldsymbol{V}^* \tag{3.2.26}$$

is lower triangular, that is each error is uncorrelated with the previous components.

Simultaneous equation models

We have a simultaneous equation model when it is not possible to order the equations of a structural VAR in a way that $\boldsymbol{\Phi}_0^*$ is triangular.

Example 3.2.7 *Consider the model*

$$\begin{cases} x_t &= \alpha_1 x_{t-1} + \alpha_2 y_t + \alpha_3 z_t + u_t \\ y_t &= \beta_1 y_{t-1} + \beta_2 z_t + \beta_3 x_{t-1} + v_t \\ z_t &= \gamma_1 z_{t-1} + \gamma_2 x_t + \gamma_3 y_{t-1} + e_t. \end{cases} \tag{3.2.27}$$

In this case there is no way we can order the equations for $\boldsymbol{\Phi}_0^*$ to be triangular. □

In this situation, $\boldsymbol{\Phi}_0^*$ may be constrained differently and may not demand acyclic dependence (DAG) and the correlation between $\boldsymbol{V}_t$ and $\boldsymbol{E}_t^*$ also constrained by defining

endogenous and exogenous components of $\boldsymbol{X}_t$. Such constraints must uniquely specify the model for the purpose of estimation.

We shall not deal with identification and estimation problems for simultaneous equation models.

3.3 Multivariate MA models

If the RHS of 3.1.6 is a finite series, $\boldsymbol{X}_t$ follows a multivariate *MA model* (MA). A MA model of sufficiently high order, in a similar way to the VAR, will adequately approximate any process. A multivariate MA process of order q, MA(q), corrected for the mean is expressed in terms of the moving average parameters $\boldsymbol{\Theta}_j$ as

$$\boldsymbol{X}_t = \boldsymbol{E}_t - \boldsymbol{\Theta}_1\boldsymbol{E}_{t-1} - \boldsymbol{\Theta}_2\boldsymbol{E}_{t-2} - \ldots - \boldsymbol{\Theta}_q\boldsymbol{E}_{t-q} = \boldsymbol{\Theta}(\mathrm{B})\boldsymbol{E}_t. \tag{3.3.1}$$

Because it is a representation of the Wold's stochastic component of the process it is always stationary. For this representation to be valid, the model must be invertible, i.e. it should be possible to represent it as

$$\boldsymbol{X}_t = \sum_{j=1}^{\infty} \boldsymbol{\Pi}_j \boldsymbol{X}_{t-j} + \boldsymbol{E}_t \tag{3.3.2}$$

that is an infinite order VAR when $\boldsymbol{E}_t$ is the linear innovation of $\boldsymbol{X}_t$. Rewrite 3.3.2 to define the operator $\boldsymbol{\Pi}(\mathrm{B})$ by

$$\boldsymbol{X}_t - \sum_{j=1}^{\infty} \boldsymbol{\Pi}_j \boldsymbol{X}_{t-j} = \boldsymbol{\Pi}(\mathrm{B})\boldsymbol{X}_t = \boldsymbol{E}_t. \tag{3.3.3}$$

Let $\det\{\boldsymbol{\Theta}(\mathrm{B})\}$ be the determinant of $\boldsymbol{\Theta}(\mathrm{B})$ and $\mathrm{Adj}\{\boldsymbol{\Theta}(\mathrm{B})\}$ be the adjoint matrix of $\boldsymbol{\Theta}(\mathrm{B})$, then

$$\boldsymbol{\Theta}(\mathrm{B})^{-1} = \frac{1}{\det\{\boldsymbol{\Theta}(\mathrm{B})\}}\mathrm{Adj}\{\boldsymbol{\Theta}(\mathrm{B})\} \tag{3.3.4}$$

and the multivariate MA(q) in 3.3.1 can be written in the inverted form as

$$\frac{1}{\det\{\boldsymbol{\Theta}(\mathrm{B})\}}\mathrm{Adj}\{\boldsymbol{\Theta}(\mathrm{B})\}\boldsymbol{X}_t = \boldsymbol{E}_t. \tag{3.3.5}$$

From 3.3.3 and 3.3.5 we have that $\{\boldsymbol{X}_t\}$ is invertible if

$$\boldsymbol{\Pi}(\mathrm{B}) = \frac{1}{\det\{\boldsymbol{\Theta}(\mathrm{B})\}}\mathrm{Adj}\{\boldsymbol{\Theta}(\mathrm{B})\} \tag{3.3.6}$$

may be expanded as a convergent series for $|\ \mathrm{B}\ | < 1$. Thus, for a multivariate MA(q) to be invertible it is required that

$$\det\{\boldsymbol{\Theta}(\mathrm{B})\} \neq 0 \quad \text{for} \quad |\ \mathrm{B}\ | \leq 1. \tag{3.3.7}$$

3.4 Multivariate ARMA models

It is natural to apply to the multivariate process, the same step of extension needed for the univariate process, which moves from the pure autoregressive and pure MA model to the mixed ARMA model. This is done simply by including both AR and MA terms in the same model giving

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \ldots + \boldsymbol{\Phi}_p \boldsymbol{X}_{t-p} + \boldsymbol{E}_t - \boldsymbol{\Theta}_1 \boldsymbol{E}_{t-1} - \ldots - \boldsymbol{\Theta}_q \boldsymbol{E}_{t-q} \tag{3.4.1}$$

This is a true extension of the AR and MA classes in the sense that an ARMA process with order $p > 0$, $q > 0$ cannot in general be represented exactly by an AR model or MA model alone. In practice an ARMA model of low order may fit well a process which could be approximated only by an AR or MA model of much higher order.

Example 3.4.1 *(Multivariate ARMA(1,1)). A multivariate ARMA(1,1) model is given by the equation system*

$$\left\{ \begin{array}{rcl} x_t &=& \phi_{1,1} x_{t-1} + \phi_{1,2} y_{t-1} + e_t - \theta_{1,1} e_{t-1} - \theta_{1,2} f_{t-1} \\ y_t &=& \phi_{2,1} x_{t-1} + \phi_{2,2} y_{t-1} + e_t - \theta_{2,1} e_{t-1} - \theta_{2,2} f_{t-1} \end{array} \right. \tag{3.4.2}$$

or, in matrix form, by

$$\left[\begin{array}{c} x_t \\ y_t \end{array} \right] = \left[\begin{array}{cc} \phi_{1,1} & \phi_{1,2} \\ \phi_{2,1} & \phi_{2,2} \end{array} \right] \left[\begin{array}{c} x_{t-1} \\ y_{t-1} \end{array} \right] + \left[\begin{array}{c} e_t \\ f_t \end{array} \right] - \left[\begin{array}{cc} \theta_{1,1} & \theta_{1,2} \\ \theta_{2,1} & \theta_{2,2} \end{array} \right] \left[\begin{array}{c} e_{t-1} \\ f_{t-1} \end{array} \right]. \tag{3.4.3}$$

□

The orders p, q of the AR and MA parts of the model are the maximum lags of $\boldsymbol{X}_t$ and $\boldsymbol{E}_t$ in 3.4.1. In practice some, or many of the elements of the coefficient matrices $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$ may be zero. In the case when $\boldsymbol{\Phi}_p$ and $\boldsymbol{\Theta}_q$ have some non-zero elements, or more generally when they are rank deficient, it is possible that the model of the given orders p,q is not unique.

Example 3.4.2 *Consider the multivariate ARMA (1,1) model*

$$(\boldsymbol{\Phi}_0 - \boldsymbol{\Phi}_1 B)\boldsymbol{X}_t = (\boldsymbol{\Theta}_0 - \boldsymbol{\Theta}_1 B)\boldsymbol{E}_t \tag{3.4.4}$$

and more specifically

$$\left\{ \boldsymbol{I} - \left[\begin{array}{cc} 0.2 & 0.4 \\ 0.1 & 0.2 \end{array} \right] B \right\} \boldsymbol{X}_t = \left\{ \boldsymbol{I} - \left[\begin{array}{cc} 0.6 & 0.8 \\ 0.4 & 0.4 \end{array} \right] B \right\} \boldsymbol{E}_t \tag{3.4.5}$$

or in equation form

$$\begin{cases} x_t - 0.2x_{t-1} - 0.4y_{t-1} &= e_t - 0.6e_{t-1} - 0.8f_{t-1} \\ x_t - 0.1x_{t-1} - 0.2y_{t-1} &= e_t - 0.3e_{t-1} - 0.4f_{t-1}. \end{cases} \tag{3.4.6}$$

Now let us suppose we have a non-zero vector $\boldsymbol{n}$ *such that, premultiplying by* $\boldsymbol{n}$ *the coefficient partitioned matrix, we obtain a zero vector, formally*

$$\boldsymbol{n}[\boldsymbol{\Phi}_p \ \boldsymbol{\Theta}_q] = 0 \tag{3.4.7}$$

i.e

$$[\ n_1 \quad n_2\ ][\ \boldsymbol{\Phi}_p \quad \boldsymbol{\Theta}_q\ ] = 0 \tag{3.4.8}$$

where n_1 *and* n_2 *are the elements of the vector* $\boldsymbol{n}$*. In our case equation 3.4.8 becomes*

$$[\ n_1 \quad n_2\ ][\ \boldsymbol{\Phi}_1 \quad \boldsymbol{\Theta}_1\ ] = 0 \tag{3.4.9}$$

or

$$[\ 1 \quad -2\ ] \begin{bmatrix} 0.2 & 0.4 & 0.6 & 0.8 \\ 0.1 & 0.2 & 0.3 & 0.4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.4.10}$$

If we premultiply both sides of 3.4.4 by $(\boldsymbol{I} + \boldsymbol{wn}B)$*, where* $\boldsymbol{w}$ *is an arbitrary vector, we obtain*

$$(\boldsymbol{I} + \boldsymbol{wn}B)(\boldsymbol{I} - \boldsymbol{\Phi}_1 B)\boldsymbol{X}_t = (\boldsymbol{I} + \boldsymbol{wn}B)(\boldsymbol{I} - \boldsymbol{\Theta}_1 B)\boldsymbol{E}_t \tag{3.4.11}$$

which, because $\boldsymbol{n}\boldsymbol{\Phi}_1 = \boldsymbol{n}\boldsymbol{\Theta}_1 = 0$*, becomes*

$$(\boldsymbol{I} + \boldsymbol{wn}B - \boldsymbol{\Phi}_1 B)\boldsymbol{X}_t = (\boldsymbol{I} + \boldsymbol{wn}B - \boldsymbol{\Theta}_1 B)\boldsymbol{E}_t. \tag{3.4.12}$$

Let $w = \begin{bmatrix} 0.5 \\ 0.7 \end{bmatrix}$ *then 3.4.12, considering also 3.4.10, becomes*

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} 0.5 \\ 0.7 \end{bmatrix} \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} - \begin{bmatrix} 0.2 & 0.4 \\ 0.1 & 0.2 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} =$$
$$= \begin{bmatrix} e_t \\ f_t \end{bmatrix} + \begin{bmatrix} 0.5 \\ 0.7 \end{bmatrix} \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} e_{t-1} \\ f_{t-1} \end{bmatrix} - \begin{bmatrix} 0.6 & 0.8 \\ 0.4 & 0.4 \end{bmatrix} \begin{bmatrix} e_{t-1} \\ f_{t-1} \end{bmatrix} \tag{3.4.13}$$

and then solving

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} 0.3 & -1.4 \\ 0.6 & -1.6 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} e_t \\ f_t \end{bmatrix} + \begin{bmatrix} -0.1 & -1.8 \\ 0.4 & -1.8 \end{bmatrix} \begin{bmatrix} e_{t-1} \\ f_{t-1} \end{bmatrix} \tag{3.4.14}$$

or, in equation form

$$\begin{cases} x_t + 0.3x_{t-1} - 1.4y_{t-1} &= e_t - 0.1e_{t-1} - 1.8f_{t-1} \\ y_t + 0.6x_{t-1} - 1.6y_{t-1} &= f_t + 0.4e_{t-1} - 1.8f_{t-1}. \end{cases} \tag{3.4.15}$$

This model still has $\mathbf{\Phi}_0 = \mathbf{\Theta}_0 = \boldsymbol{I}$*, order* $p = 1$ *and* $q = 1$*; nevertheless 3.4.15 is different from 3.4.4, although it is exactly equivalent and must represent the same stochastic structure. This means that even if the (minimum) order p,q of the ARMA model is correctly identified, it is possible that the model is not unique. This must be a consideration when fitting models to data.* □

If it is known that the model is not unique in the way shown in the example, constraints can be imposed upon $\mathbf{\Phi}_p$ and $\mathbf{\Theta}_p$ to obtain a unique representation by the ARMA(p,q) (see Reinsel, 1993, pp. 36–39).

3.5 Identification of multivariate time series models

Several different approaches have been proposed to identify the order of multivariate time series models. We discuss in this chapter the extensions of methods already seen for the univariate case. In particular we discuss the use of the sample ACF to identify the multivariate MA model, the Yule-Walker equations for the VAR and the extended Yule-Walker equations for the order identification of multivariate ARMA models.

3.5.1 Multivariate MA models

The cross-covariance matrix, at lag l, of a multivariate MA(q) process is given by

$$\begin{aligned} \text{Cov}\,(\boldsymbol{X}_t, \boldsymbol{X}_{t-l}) &= \mathbf{\Gamma}(l) \\ &= \text{E}[(\boldsymbol{E}_t - \mathbf{\Theta}_1 \boldsymbol{E}_{t-1} - \ldots - \mathbf{\Theta}_q \boldsymbol{E}_{t-q}) \\ &\quad (\boldsymbol{E}_{t-l} - \mathbf{\Theta}_1 \boldsymbol{E}_{t-1-l} - \ldots - \mathbf{\Theta}_q \boldsymbol{E}_{t-q-l})]. \end{aligned} \tag{3.5.1}$$

If $\mid l \mid \leq \mid q \mid$, then

$$\begin{aligned} \mathbf{\Gamma}(l) &= -\mathbf{\Sigma}\mathbf{\Theta}_l^T + \mathbf{\Theta}_1 \mathbf{\Sigma}\mathbf{\Theta}_{l-1}^T + \ldots + \mathbf{\Theta}_m \mathbf{\Sigma}\mathbf{\Theta}_q^T \\ &= \sum_{h=0}^{m} \mathbf{\Theta}_h \mathbf{\Sigma}\mathbf{\Theta}_{l-h}^T \end{aligned} \tag{3.5.2}$$

where $l - m = q$ and $\mathbf{\Theta}_0 = -\boldsymbol{I}$. Because of error independence it is apparent that

$$\mathbf{\Gamma}(l) = 0 \qquad \text{for} \qquad l > q. \tag{3.5.3}$$

This property, as seen for the univariate case, turns to be very useful to identify the order of a multivariate MA model. Because of 3.5.3 also all the cross-correlations are zero for

lags greater than q. Hence to identify the order we can observe the maximum lag, q, for which at least one element of $\mathbf{\Gamma}(\mathrm{q})$ is significantly different from zero. In reality we can observe only a sample covariance matrix $\hat{\mathbf{\Gamma}}(\mathrm{q})$ whose elements, $\hat{\gamma}_{ij}(l)$, consistently with 2.7.1, are given by

$$\hat{\gamma}_{ij}(l) = \frac{\sum_{t=1}^{n-l}(x_{i,t} - \hat{\mu}_{xi})(x_{j,t-l} - \hat{\mu}_{xj})}{n} \tag{3.5.4}$$

where $\hat{\mu}_{xi}$ and $\hat{\mu}_{xj}$ are the mean values for $x_{i,t}$ and $x_{j,t}$. In practice it is more common to observe the sample cross-correlation matrix $\hat{\mathbf{\Pi}}(l)$ whose elements, $\hat{\rho}_{ij}(l)$ are given by

$$\hat{\rho}_{ij}(l) = \frac{\hat{\gamma}_{ij}(l)}{\sqrt{\hat{\gamma}_{ii}(0)\hat{\gamma}_{jj}(0)}} \tag{3.5.5}$$

Unfortunately the sampling properties of $\hat{\rho}_{ij}(l)$ are complicated and depend on the unknown theoretical values of $\rho_{ij}(l)$. For a stationary vector process, when the number of observations is large, the sample values $\hat{\rho}_{ij}(l)$ are asymptotically normally distributed with corresponding means $\rho_{ij}(l)$. The variances and covariances of the estimates are complicated. For a Gaussian process (see Reinsel, 1993, p. 76), we have

$$\begin{aligned}\mathrm{Cov}\ (\hat{\rho}_{ij}(l), \hat{\rho}_{ij}(n)) \approx \frac{1}{n}\sum_{u=-\infty}^{\infty}\Big\{ & \rho_{ii}(u)\rho_{jj}(u+n-l) + \rho_{ij}(u+n)\rho_{ji}(u-l) \\ - \ & \rho_{ij}(l)\left[\rho_{ii}(u)\rho_{ij}(u+n) + \rho_{jj}(u)\rho_{ji}(u-n)\right] \\ - \ & \rho_{ij}(n)\left[\rho_{ii}(u)\rho_{ij}(u+l) + \rho_{jj}(u)\rho_{ji}(u-l)\right] \\ + \ & \rho_{ij}(l)\rho_{ij}(n)\left[\frac{1}{2}\rho_{ii}^2(u) + \rho_{ij}^2(u) + \frac{1}{2}\rho_{jj}^2(u)\right]\Big\}.\end{aligned} \tag{3.5.6}$$

Setting $n = l$ we obtain an expression for the asymptotic variance of $\hat{\rho}_{ij}(l)$. We can derive simplified variances for some special cases.

Example 3.5.1 *(White noise). Suppose that $\{\boldsymbol{X}_t\}$ is a vector white noise process, with covariance matrix $\mathbf{\Sigma}$ and correlation matrix $\mathbf{\Pi}(0)$, so that $\pi_{ij}(l) = 0$ for $l \neq 0$, then 3.5.6 yields*

$$\mathrm{Var}\,(\hat{\rho}_{ij}(l)) \approx \frac{1}{n} \qquad l \neq 0 \tag{3.5.7}$$

$$\mathrm{Var}\,(\hat{\rho}_{ij}(0)) \approx \frac{1}{n}\left[1 - \rho_{ij}^2(0)\right]^2 \qquad i \neq j \tag{3.5.8}$$

$$\mathrm{Cov}\,(\hat{\rho}_{ij}(l), \hat{\rho}_{ij}(-l)) \approx \frac{1}{n}\,\hat{\rho}_{ij}^2(0). \tag{3.5.9}$$

Hence from 3.5.9 and 3.5.7 we have

$$\mathrm{Corr}\,(\hat{\rho}_{ij}(l), \hat{\rho}_{ji}(l)) \approx \rho_{ij}^2(0) \tag{3.5.10}$$

In any other case

$$\text{Cov}\left(\hat{\rho}_{ij}(l), \hat{\rho}_{ij}(n)\right) \approx 0. \tag{3.5.11}$$

□

Example 3.5.2 *(MA(1)). Suppose we have a MA(1) process, so that* $\rho_{ij}(l) = 0$ *for* $l > 1$*. Hence in this case from 3.5.6 we have*

$$\text{Var}\left(\hat{\rho}_{ii}(1)\right) \approx \frac{1}{n}\left[1 - 3\rho_{ii}^2(1) + 4\rho_{ii}^4(1)\right] \tag{3.5.12}$$

and

$$\text{Var}\left(\hat{\rho}_{ij}(l)\right) \approx \frac{1}{n}\left[1 + 2\rho_{ii}(1)\rho_{jj}(1)\right] \quad \textit{for} \quad l = \pm 2, \pm 3, \ldots \tag{3.5.13}$$

□

Example 3.5.3 *(MA(q)). Similarly to the previous example, if* $\{\boldsymbol{X}_t\}$ *is a multivariate MA(q) process, so that* $\rho_{ij}(l) = 0$ *for* $l > q$*, then we have*

$$E[\hat{\rho}_{ij}(l)] \approx 0 \tag{3.5.14}$$

and

$$\text{Var}\left(\hat{\rho}_{ij}(l)\right) \approx \frac{1}{n}\left[1 + 2\sum_{u=1}^{q}\rho_{ii}(u)\rho_{jj}(u)\right] \quad \textit{for} \quad |\, l \,| > q. \tag{3.5.15}$$

□

Results as in 3.5.13 can be used to check the significance of individual sample cross-correlations $\hat{\rho}_{ij}(l)$ for $l > 1$ in assessing the appropriateness of a vector MA(1) model for the series $\boldsymbol{X}_t$. In the same way results from 3.5.15 can be used to check the significance of individual $\hat{\rho}_{ij}(l)$ for $l > q$ when considering a low order MA(q) model for $\boldsymbol{X}_t$. In practice expression 3.5.15 is used with the unknown $\rho_{ii}(u)$ and $\rho_{jj}(u)$ replaced by the estimates $\hat{\rho}_{ii}(u)$ and $\hat{\rho}_{jj}(u)$. Tiao and Box (1981) proposed as a rough guideline a significance interval of $\pm\frac{2}{\sqrt{n}}$, where $\pm\frac{1}{\sqrt{n}}$ is the standard deviation appropriate for the white noise case as derived in 3.5.7.

In the same paper Tiao and Box proposed also a way to deal with the large number of correlations to examine by displaying the results in matrices. Nevertheless, this still remains a problem when we consider a high order for the multivariate MA model.

3.5.2 Multivariate AR models

To identify multivariate AR models we extend the univariate concept of partial autocorrelation. Again, partial correlations measure the contribution to a regression of a new term, which is additional to the contribution of existing terms. A non-zero partial correlation indicates that the new term should be included in the regression. Conversely, zero partial correlations indicate that the new term is not necessary, so that the existing terms are sufficient. In the context of VAR models this is useful to determine when the order is sufficient.

An extension to the multivariate context of the *Durbin-Levinson* method of solving the *Yule-Walker equations* of AR models has been presented by Whittle (1963). This provides a useful and efficient way of evaluating partial correlations and the autoregressive coefficients in a VAR model. An important point is that for multivariate processes we have to consider a forward autoregression and a backward autoregression with different coefficients. Consider first the regression

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \ldots + \boldsymbol{\Phi}_p \boldsymbol{X}_{t-p} + \boldsymbol{E}_t \tag{3.5.16}$$

the parameters of which are determined by the requirement that $\boldsymbol{E}_t \perp \boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-p}$. This gives the *multivariate Yule-Walker equations*

$$\begin{array}{ccccccc} \boldsymbol{\Gamma}_1 & = & \boldsymbol{\Phi}_1\boldsymbol{\Gamma}_0 & + & \ldots & + & \boldsymbol{\Phi}_p\boldsymbol{\Gamma}_{p-1} \\ \vdots & & \vdots & & \ldots & & \vdots \\ \boldsymbol{\Gamma}_p & = & \boldsymbol{\Phi}_1\boldsymbol{\Gamma}_{p-1} & + & \ldots & + & \boldsymbol{\Phi}_p\boldsymbol{\Gamma}_0 \end{array} \tag{3.5.17}$$

and the error variance matrix is

$$\boldsymbol{V} = \boldsymbol{\Gamma}_0 - \boldsymbol{\Phi}_1\boldsymbol{\Gamma}_1 - \ldots - \boldsymbol{\Phi}_p\boldsymbol{\Gamma}_p. \tag{3.5.18}$$

Although equations like 3.5.16 are called autoregressions, they are not strictly AR models, unless the order p is sufficiently high that $\boldsymbol{E}_t$ is a multivariate white noise. This is the order we want to identify. To follow the univariate approach, we also need to represent $\boldsymbol{X}_{t-p}$ by a reverse time autoregression

$$\boldsymbol{X}_{t-p-1} = \boldsymbol{\Phi}_1^* \boldsymbol{X}_{t-p} + \ldots + \boldsymbol{\Phi}_p^* \boldsymbol{X}_{t-1} + \boldsymbol{E}_{t-p-1}^* \tag{3.5.19}$$

where $\boldsymbol{E}_{t-p-1}^* \perp \boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-p}$. If then, $\boldsymbol{E}_t \perp \boldsymbol{E}_{t-p-1}^*$ the forward regression 3.5.16 cannot be improved by extending to include $\boldsymbol{X}_{t-p-1}$. Thus

$$\mathrm{Cov}\,(\boldsymbol{E}_t, \boldsymbol{E}_{t-p-1}^*) = \boldsymbol{Q}_{p+1} \tag{3.5.20}$$

is the block partial covariance between $\boldsymbol{X}_t$ and $\boldsymbol{X}_{t-p-1}$. It is more practicable to assess correlation than covariance, so we consider correlation matrices derived from $\boldsymbol{Q}_{p+1}$. Let $\mathrm{Var}\,(\boldsymbol{E}_t) = \boldsymbol{V}_{p+1}$ and $\mathrm{Var}\,(\boldsymbol{E}^*_{t-p-1}) = \boldsymbol{V}^*_{p+1}$ then there are two main possible definitions of the partial correlation matrix. One is to consider a matrix whose elements are

$$\mathrm{Corr}\,(\boldsymbol{E}_{t,i}, \boldsymbol{E}^*_{t-p-1,j}) = \frac{(\boldsymbol{Q}_{p+1})_{i,j}}{\sqrt{(\boldsymbol{V}_{p+1})_{i,i}}\sqrt{(\boldsymbol{V}^*_{p+1})_{j,j}}} \tag{3.5.21}$$

where $(\boldsymbol{Q}_{p+1})_{i,j}$ and $(\boldsymbol{V}_{p+1})_{j,j}$ are respectively elements of $\boldsymbol{Q}_{p+1}$ and $\boldsymbol{V}_{p+1}$ (see Reinsel, 1993,p. 69–70).

Another approach, proposed by Ansley and Newbold (1979), is to consider the relation

$$\boldsymbol{P}_{p+1} = \boldsymbol{V}^{-\frac{1}{2}}_{p+1}\boldsymbol{Q}_{p+1}\boldsymbol{V}^{*-\frac{1}{2}}_{p+1} \tag{3.5.22}$$

where $\boldsymbol{V}^{-\frac{1}{2}}_{p+1}$ is the symmetric square root matrix of $\boldsymbol{V}_{p+1}$. That is one possible way based upon standardising $\boldsymbol{E}_t$ and $\boldsymbol{E}^*_t$ to unit variance matrices. The standard way of doing this is to apply the Choleski factorisation to $\boldsymbol{V}_{p+1}$ so that

$$\boldsymbol{V}_{p+1} = \boldsymbol{T}_{p+1}\boldsymbol{T}^T_{p+1} \tag{3.5.23}$$

where $\boldsymbol{T}_{p+1}$ is the lower triangular Choleski factor. Then let

$$\boldsymbol{F}_t = \boldsymbol{T}^{-1}_{p+1}\boldsymbol{E}_t. \tag{3.5.24}$$

Hence we have

$$\mathrm{Var}\,(\boldsymbol{F}_t) = \boldsymbol{I} = \mathrm{Var}\,(\boldsymbol{F}^*_{t-p-1}) \tag{3.5.25}$$

and

$$\mathrm{E}(\boldsymbol{F}_t\boldsymbol{F}^*_{t-p-1}) = \boldsymbol{P}_{p+1} = \boldsymbol{T}^{-1}_{p+1}\boldsymbol{Q}_{p+1}\boldsymbol{T}^{-T}_{p+1} \tag{3.5.26}$$

which is a matrix of correlations between $\boldsymbol{F}_t$ and $\boldsymbol{F}^*_{t-p-1}$.

In this case, where the process dimension is k, the last element of $\boldsymbol{F}_t$, $\boldsymbol{F}_{t,k}$, is the error in regression of $\boldsymbol{E}_{t,k}$ on $\boldsymbol{E}_{t,1}, \ldots, \boldsymbol{E}_{t,k-1}$ and the correlation must be interpreted as

$$(\boldsymbol{P}_{p+1})_{k,k} = \mathrm{Corr}\,(X_{t,k}, X_{t-p-1,k} \mid \{\boldsymbol{X}_t, \ldots, \boldsymbol{X}_{t-p-1}\} \setminus \{X_{t,k}, X_{t-p-1,k}\}) \tag{3.5.27}$$

which is the true scalar partial correlation between $\boldsymbol{X}_{t,k}$ and $\boldsymbol{X}_{t-p-1,k}$ in the whole set $\boldsymbol{X}_t, \ldots, \boldsymbol{X}_{t-p-1}$.

3.5.3 Multivariate ARMA models

Similarly to the univariate case we approach the identification of multivariate ARMA models by the extended Yule-Walker equations.

From Wold's decomposition we know that we can represent a purely random mean corrected process $\{\boldsymbol{X}_{t-l}\}$ by the infinite MA process

$$\boldsymbol{X}_{t-l} = \sum_{i=0}^{\infty} \boldsymbol{\Psi}_i \boldsymbol{E}_{t-l-i}$$

and then

$$\mathrm{E}(\boldsymbol{X}_{t-l}\boldsymbol{E}_{t-l}) = \boldsymbol{\Psi}_{j-l}\boldsymbol{V}. \tag{3.5.28}$$

Thus considering the vector ARMA representation as in 3.4.1

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_1\boldsymbol{X}_{t-1} + \ldots + \boldsymbol{\Phi}_p\boldsymbol{X}_{t-p} + \boldsymbol{E}_t - \boldsymbol{\Theta}_1\boldsymbol{E}_{t-1} - \ldots - \boldsymbol{\Theta}_q\boldsymbol{E}_{t-q}$$

it is easy to determine that its covariance matrix satisfies the relation

$$\mathrm{Cov}\,(\boldsymbol{X}_{t-l}, \boldsymbol{X}_t) = \sum_{j=1}^{p} \mathrm{Cov}\,(\boldsymbol{X}_{t-l}, \boldsymbol{X}_{t-j})\boldsymbol{\Phi}_j^T + \mathrm{Cov}\,(\boldsymbol{X}_{t-l}, \boldsymbol{E}_t) - \sum_{j=1}^{q} \mathrm{Cov}\,(\boldsymbol{X}_{t-l}, \boldsymbol{E}_{t-j})\boldsymbol{\Theta}_j^T \tag{3.5.29}$$

and hence

$$\boldsymbol{\Gamma}(l) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(l-j)\boldsymbol{\Phi}_j^T - \sum_{j=l}^{q} \boldsymbol{\Psi}_{j-l}\boldsymbol{V}\boldsymbol{\Theta}_j^T, l = 0, 1, \ldots, q \tag{3.5.30}$$

with the convention that $\boldsymbol{\Theta}_0 = \boldsymbol{I}$.

If $l > q$, because of the non-correlation of the residuals, 3.5.30 becomes

$$\boldsymbol{\Gamma}(l) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(l-j)\boldsymbol{\Phi}_j^T. \tag{3.5.31}$$

For $l = q+1, \ldots, q+p$ the last equation provides a set of linear equations which gives a unique solution for the coefficient matrices $\boldsymbol{X}_t = \boldsymbol{\Phi}_j$ in terms of $\boldsymbol{\Gamma}(l)$ provided that $\boldsymbol{\Gamma}(l-j)$ is of full rank.

In practice to identify a multivariate ARMA model, once a sufficiently large value of p is chosen, one should estimate the parameters of the pure autoregressive component by 3.5.31 and then check the autocorrelation function of the residual

$$\boldsymbol{X}_t - \boldsymbol{\Phi}_1\boldsymbol{X}_{t-1} + \ldots + \boldsymbol{\Phi}_p\boldsymbol{X}_{t-p}$$

for an increasing value of p. While there is still an AR component left in the residual, the ACF will decrease slowly without any sudden cut. Once we have specified all the AR components the ACF of the residual will assume the characteristic shape of a MA process, thus revealing the order of both the AR and MA components.

3.6 Estimation of multivariate time series models

In this paragraph we discuss how to estimate multivariate ARMA models. In particular we begin discussing the Yule-Walker equations approach and the least squares method. Then we treat the likelihood approach for multivariate AR, MA and ARMA.

3.6.1 Multivariate AR models

The estimation of parameters for a multivariate AR model by the Yule-Walker equations is, like the univariate case, straightforward. Equations 3.5.17 can be written in block matrix form as

$$\boldsymbol{\Gamma}_{(p)} = \boldsymbol{\Phi}^{\star}\boldsymbol{\Gamma}_{p} \tag{3.6.1}$$

where

$$\boldsymbol{\Phi}^{\star} = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_p), \qquad \boldsymbol{\Gamma}_{(p)} = (\boldsymbol{\Gamma}(1), \boldsymbol{\Gamma}(2), \ldots, \boldsymbol{\Gamma}(p))$$

and $\boldsymbol{\Gamma}_p$ is a $kp \times kp$ matrix with (i,j)th block of elements equal to $\boldsymbol{\Gamma}(i-j)$. Hence, from 3.6.1 we have

$$\boldsymbol{\Phi}^{\star} = \boldsymbol{\Gamma}_{(p)}\boldsymbol{\Gamma}_{p}^{-1} \tag{3.6.2}$$

Alternatively we can estimate the parameters in a VAR model by the least squares method. If $\boldsymbol{X}_t$ follows a mean corrected AR(p) model we can rearrange the equations for $t = 1+p, \ldots, k$ as for the classical linear model. Consider the separate regressions of each series upon lagged values of all the others. For series j, let the response vector, regressors and errors be

$$\boldsymbol{Y}_j = \begin{bmatrix} X_{p+1,j} \\ \vdots \\ \vdots \\ X_{n,j} \end{bmatrix};$$

$$\boldsymbol{X}_j = \begin{bmatrix} X_{p,1} & \ldots & X_{p,k} & X_{p-1,1} & \ldots & X_{p-1,k} & \ldots & X_{1,1} & \ldots & X_{1,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{n-1,1} & \ldots & X_{n-1,k} & X_{n-2,1} & \ldots & X_{n-2,k} & \ldots & X_{n-p,1} & \ldots & X_{n-p,k} \end{bmatrix};$$

$$\boldsymbol{E}_j = \begin{bmatrix} E_{p+1,j} \\ \vdots \\ \vdots \\ E_{n,j} \end{bmatrix}$$

with parameters

$$\boldsymbol{\Phi}_j = (\Phi_{j,11}, \ldots, \Phi_{j,k1}, \Phi_{j,12}, \ldots, \Phi_{j,k2}, \ldots, \Phi_{j,1p}, \ldots, \Phi_{j,kp}).$$

Note that the regression matrix $\boldsymbol{X}$ is the same for each response $\boldsymbol{Y}_j$. The set of least squares equations is

$$(\boldsymbol{X}_j^T \boldsymbol{X}_j)\boldsymbol{\Phi}_j = \boldsymbol{X}_j^T \boldsymbol{Y}_j. \tag{3.6.3}$$

The elements of $\frac{1}{k-p}\boldsymbol{X}^T\boldsymbol{X}$ are of the form

$$\frac{1}{k-p}\sum_{t=1+p}^{k} \boldsymbol{X}_{i,t-k}\boldsymbol{X}_{j,t-l}. \tag{3.6.4}$$

These are estimates of $\boldsymbol{\Gamma}_{ij}(k-l)$ which if replaced by the standard estimates $\hat{\boldsymbol{\Gamma}}_{ij}(k-l)$ give the Yule-Walker equations. They differ from the standard estimates in the way the series cut off at the ends.

The k regressions, $\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{\Phi}_j + \boldsymbol{E}_j$, minimise separately, the sums of squares of errors

$$S_1 = \sum_{t=1+p}^{k} e_{1,t}^2, \ldots, S_k = \sum_{t=1+p}^{k} e_{k,t}^2. \tag{3.6.5}$$

These are seemingly unrelated regressions, but the errors in one regression are contemporaneously correlated with errors in the other regressions.

A fully efficient way to estimate the parameters of a VAR model is by maximum likelihood (Judge *et al.*, 1985, p. 468). Similarly to the methods used for the univariate case we shall consider the simple case of a VAR(1). In this case the likelihood is given by

$$L(\boldsymbol{\Phi}) \propto f(\boldsymbol{X}_1)\prod_{t=2}^{n} f(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}) \tag{3.6.6}$$

$$\propto \frac{1}{\mid \boldsymbol{\Gamma}(0) \mid^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\boldsymbol{X}_1^T\boldsymbol{\Gamma}(0)^{-1}\boldsymbol{X}_1\right\} \times \prod_{t=2}^{n}\frac{1}{\mid \boldsymbol{V}^{\frac{1}{2}} \mid}\exp\left\{-\frac{1}{2}(\boldsymbol{X}_t - \boldsymbol{\Phi}\boldsymbol{X}_{t-1})^T\boldsymbol{V}^{-1}(\boldsymbol{X}_t - \boldsymbol{\Phi}\boldsymbol{X}_{t-1})\right\}. \tag{3.6.7}$$

One approach to maximum likelihood is to treat $\boldsymbol{X}_1$ as a fixed value and thus obtaining the maximum likelihood estimate conditioning on $\boldsymbol{X}_1$ and minimising

$$n^* \log(\mid \boldsymbol{V} \mid) + \sum_{t=2}^{n}(\boldsymbol{X}_t - \boldsymbol{\Phi}\boldsymbol{X}_{t-1})^T\boldsymbol{V}^{-1}(\boldsymbol{X}_t - \boldsymbol{\Phi}\boldsymbol{X}_{t-1}) \tag{3.6.8}$$

where $n^* = n - 1$.

The parameters are the elements of $\boldsymbol{\Phi}$ and $\boldsymbol{V}$ and their number is $k^2 + \frac{1}{2}k(k+1)$. Therefore to go from the *canonical* to the *structural* form of the VAR, let $\boldsymbol{T}$ be a lower triangular matrix such that

$$\boldsymbol{T}\boldsymbol{V}\boldsymbol{T}^T = \boldsymbol{D} \tag{3.6.9}$$

where $\boldsymbol{D}$ is a diagonal matrix whose non-zero elements are the variances σ_j^2. Hence from 3.6.9 we have

$$\boldsymbol{V}^{-1} = \boldsymbol{T}^T\boldsymbol{D}^{-1}\boldsymbol{T}. \tag{3.6.10}$$

Moreover,

$$|\ \boldsymbol{V}\ | = |\ \boldsymbol{D}\ | \tag{3.6.11}$$

and

$$\log(|\ \boldsymbol{V}\ |) = \sum_{i=1}^{k} \log(\sigma_i^2). \tag{3.6.12}$$

Then

$$\begin{aligned} (\boldsymbol{X}_t - \boldsymbol{\Phi}\boldsymbol{X}_{t-1})^T\boldsymbol{V}^{-1}(\boldsymbol{X}_t - \boldsymbol{\Phi}\boldsymbol{X}_{t-1}) &= (\boldsymbol{T}\boldsymbol{X}_t - \boldsymbol{T}\boldsymbol{\Phi}\boldsymbol{X}_{t-1})^T\boldsymbol{D}^{-1}(\boldsymbol{T}\boldsymbol{X}_t - \boldsymbol{T}\boldsymbol{\Phi}\boldsymbol{X}_{t-1}) \\ &= (\boldsymbol{T}\boldsymbol{X}_t - \boldsymbol{\Phi}^*\boldsymbol{X}_{t-1})^T\boldsymbol{D}^{-1}(\boldsymbol{T}\boldsymbol{X}_t - \boldsymbol{\Phi}^*\boldsymbol{X}_{t-1}). \end{aligned} \tag{3.6.13}$$

Thus the transformation from $\boldsymbol{\Phi}$ and $\boldsymbol{V}$ to $\boldsymbol{\Phi}^* = \boldsymbol{T}\boldsymbol{\Phi}$, $\boldsymbol{T}$ and $\boldsymbol{D}$ is a one to one function. Then the likelihood function becomes

$$\sum_{i=1}^{k}\left[n^*\log\left(\sigma_i^2\right) + \frac{1}{\sigma_i^2}\sum_{t=2}^{n}\epsilon_{i,t}^2\right] \tag{3.6.14}$$

where

$$\epsilon_{i,t} = X_{it} - \sum_{j=1}^{i-1} T_{i,j}X_{j,t} - \sum_{j=1}^{k}\phi_{i,j}^* X_{j,t-1} \tag{3.6.15}$$

that is, $\epsilon_{i,t}$ is the residual of the regression of $X_{i,t}$ on *previous* contemporaneous values of $X_{j,t}$ (in the ordering $X_{1,t}, \ldots, X_{k,t}$) and lagged values $X_{j,t-1}$. These are k truly separate regressions with separate parameters in each of $\left\{n^*\log\left(\sigma_i^2\right) + \frac{1}{\sigma_i^2}\sum_{t=2}^{n}\epsilon_{i,t}^2\right\}$ and $\epsilon_{i,t}$, $\epsilon_{j,t}$ are uncorrelated. For $i = 1$ the regression 3.6.15 corresponds exactly to regression 3.6.3, because for $i = 1$ there is no regression on contemporaneous variables $X_{j,t}$ and the first row of $\boldsymbol{\Phi}^*$ is the same as the first row of $\boldsymbol{\Phi}$. So regression estimates for the first row of $\boldsymbol{\Phi}$ are the same as maximum likelihood (ML) estimates. We can order the equations so that each can be first and hence, eventually, for every row we can obtain ML estimates

by regression. The one to one correspondence of the transformation is possible only because the lagged regressors are the same for each element of $X_{i,t}$; for example, subset regression of canonical autoregressions cannot be estimated in this way.

In 3.6.6, alternatively, we can incorporate the term in $\boldsymbol{X}_1$. We need to evaluate $\boldsymbol{\Gamma}_0$ as a function of $\boldsymbol{V}$ and $\boldsymbol{\Phi}$; from $\boldsymbol{X}_t = \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{E}_t$ we have

$$\boldsymbol{\Gamma}(0) = \boldsymbol{\Phi}\boldsymbol{\Gamma}(0)\boldsymbol{\Phi}^T + \boldsymbol{V} \tag{3.6.16}$$

which is a *Lyapunov equation* in terms of $\boldsymbol{\Phi}$ and $\boldsymbol{V}$ which can be solved for $\boldsymbol{\Gamma}(0)$. The term involving $\boldsymbol{X}_1$ in 3.6.7 can be evaluated and, then, using numerical optimisation a full ML estimate (MLE) can be found.

3.6.2 Multivariate MA models

Several authors have dealt with ML estimation of vector MA models (Tunnicliffe Wilson, 1973; Osborn, 1977; Phadke and Kedem, 1978; Hillmer and Tiao, 1979; Nicholls and Hall, 1979; Monti, 1998). To explain it let us consider the simple case of a MA(1)

$$\boldsymbol{X}_t = \boldsymbol{E}_t + \boldsymbol{\Theta}\boldsymbol{E}_{t-1}$$

In this case maximum likelihood estimation based on the joint distribution function of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ conditioning on $\boldsymbol{E}_0$, assuming a known value, e.g. 0 is given by minimising

$$n \log \mid \boldsymbol{V} \mid + \sum_{t=1}^{n} \boldsymbol{E}_t^T \boldsymbol{V}^{-1} \boldsymbol{E}_t. \tag{3.6.17}$$

Iterative minimisation of $\boldsymbol{V}$ can then be applied. It is possible again, at each iteration, to separate the minimisation step into k separate regressions. The iterative procedure applies also to VARMA(1,1) model, so we defer its presentation to the next section.

3.6.3 Multivariate ARMA models

We present the estimation of the multivariate ARMA(1,1) as an illustration of the general case.

Consider the model

$$\boldsymbol{X}_t = \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{E}_t - \boldsymbol{\Theta}\boldsymbol{E}_{t-1}.$$

Condition on $\boldsymbol{X}_1$ and on assumed value of $\boldsymbol{E}_1 = 0$ then for any proposed parameter $\boldsymbol{\Phi}, \boldsymbol{\Theta}$ we can regenerate, for $t = 2, \ldots, n$

$$\boldsymbol{E}_t = \boldsymbol{X}_t - \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{\Theta}\boldsymbol{E}_{t-1}. \tag{3.6.18}$$

If we assume $\boldsymbol{E}_t \sim$independent MVN$(0, \boldsymbol{V})$ the maximum likelihood is equivalent to the minimisation of the deviance

$$n^* \log \mid \boldsymbol{V} \mid + \sum_{t=2}^{n} \boldsymbol{E}_t^T \boldsymbol{V}^{-1} \boldsymbol{E}_t = n^* \log \mid \boldsymbol{V} \mid + \operatorname{tr} \left\{ \boldsymbol{V}^{-1} \sum_{t=2}^{n} \boldsymbol{E}_t \boldsymbol{E}_t^T \right\}. \tag{3.6.19}$$

Although $\boldsymbol{E}_t$ is linear in $\boldsymbol{\Phi}$, it is not linear in $\boldsymbol{\Theta}$, but a local linear approximation can be found. In terms of initial parameters $\boldsymbol{\Phi}_0$, $\boldsymbol{\Theta}_0$ and corresponding $\boldsymbol{E}_{0,t}$, consider the change in $\boldsymbol{E}_t$ following a small change in parameters $\delta\boldsymbol{\Phi} = \boldsymbol{\Phi} - \boldsymbol{\Phi}_0$, $\delta\boldsymbol{\Theta} = \boldsymbol{\Theta} - \boldsymbol{\Theta}_0$. We need to calculate the partial derivatives of $\boldsymbol{E}_t$ with respect to $\boldsymbol{\Phi}$, which are obtained by differentiating the model 3.6.18

$$\boldsymbol{E}_{\boldsymbol{\Phi},t} = -\boldsymbol{X}_{t-1} + \boldsymbol{\Theta}\boldsymbol{E}_{\boldsymbol{\Phi},t-1} \tag{3.6.20}$$

and with respect to $\boldsymbol{\Theta}$

$$\boldsymbol{E}_{\boldsymbol{\Theta},t} = \boldsymbol{E}_{t-1} + \boldsymbol{\Theta}\boldsymbol{E}_{\boldsymbol{\Theta},t-1}. \tag{3.6.21}$$

Both these series can be generated for $t = 2, \ldots, n$. Formally

$$\begin{aligned} \boldsymbol{E}_{\boldsymbol{\Phi},t} &= -(1 - \boldsymbol{\Theta}\mathrm{B})^{-1}\boldsymbol{X}_{t-1} \\ \boldsymbol{E}_{\boldsymbol{\Theta},t} &= (1 - \boldsymbol{\Theta}\mathrm{B})^{-1}\boldsymbol{E}_{t-1}. \end{aligned} \tag{3.6.22}$$

Then

$$\boldsymbol{E}_t \approx \boldsymbol{E}_{0,t} + \delta\boldsymbol{\Phi}\boldsymbol{E}_{\boldsymbol{\Phi},t} + \delta\boldsymbol{\Theta}\boldsymbol{E}_{\boldsymbol{\Theta},t} \tag{3.6.23}$$

or

$$\boldsymbol{E}_{0,t} \approx -\delta\boldsymbol{\Phi}\boldsymbol{E}_{\boldsymbol{\Phi},t} - \delta\boldsymbol{\Theta}\boldsymbol{E}_{\boldsymbol{\Theta},t} + \boldsymbol{E}_t. \tag{3.6.24}$$

Taking $\boldsymbol{E}_{0,t}$, $\boldsymbol{E}_{\boldsymbol{\Theta},t}$ and $\boldsymbol{E}_{\boldsymbol{\Phi},t}$ as known, this is a multivariate regression for $\delta\boldsymbol{\Phi}$, $\delta\boldsymbol{\Theta}$ with errors $\boldsymbol{E}_t$. Provided $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ are full matrices these can be fitted, as for the VAR(p) model, by separate regressions. New parameter values are then found as

$$\boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_0 + \delta\hat{\boldsymbol{\Phi}}, \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_0 + \delta\hat{\boldsymbol{\Theta}} \tag{3.6.25}$$

and the process iterated to convergence. The regression step of the final iteration also gives the standard errors of parameter estimates. In practice a strategy for ensuring

convergence is advisable by limiting the step size at each iteration. The strategy proposed by Marquardt (1963) can be extended to this case.

Recent papers on ML estimation of multivariate ARMA models are from: Ansley and Kohn, 1983; Reinsel *et al.*, 1992; Luceño, 1994; Mauricio, 1995.

3.7 Two examples

As practical illustrations of identification and estimation techniques as applied to multivariate time series models, in this section we illustrate the relations among three different interest rates of the Italian monetary markets and the relations among seven different maturities of U.S. dollar interest rate. Such relations are known as term structure.

3.7.1 Italian monetary market interest rates

In this subsection we take into consideration the relations between three different interest rates of the Italian monetary market: the repurchase agreement (REPO) interest rate; the treasury bond interest rate and the loan interest rate. Their time series are showed in fig. 3.1 and consist of 96 monthly observations, from January 1986 to December 1993; the source is the Bank of Italy.

As we shall see later the relation among these interest rates is important for detecting the existence of a monetary transmission mechanism known as *lending channel*. A first look at the series suggests stationarity for all of them.

Figure 3.2 shows high values for the sample autocorrelations and cross-correlations of these series up to moderate lag. Hence instead of trying to find a cut-off lag we first verify the existence of an AR component which could explain such behaviour. To do that we need to examine the partial autocorrelations (fig. 3.3).

These appear to have no significant value beyond lag 2, which implies that a VAR(2) model should be sufficient to represent the data. The coefficients and the t-values for this model are shown in table 3.1

This table shows both the Yule-Walker estimates, derived in the construction of the partial autocorrelation function, and the ordinary least squares estimates. The *end effects* present in the estimation by Yule-Walker equations constrain the estimates making, in

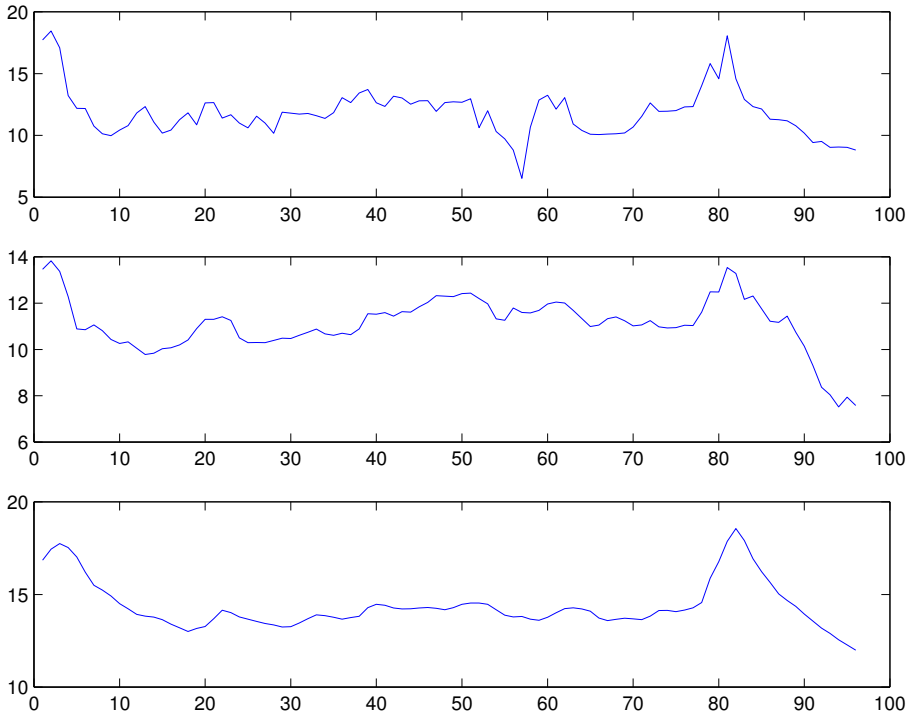


Figure 3.1: Italian monetary market interest rates; the first interest rate on the top is the repurchase agreement interest rate; the second, in the middle, is the Treasury bonds interest rate; the last, below, is the bank loan interest rate.

general, the t-values less significant. Many of the coefficients are not significant and in Chapter 5 we consider how to construct a parsimonious model to fit this dataset.

The identification of a VAR(2) is supported by the observation of the correlations and partial correlation of the residuals obtained from least square estimation (respectively figures 3.4 and 3.5) from which it appears to be adequate.

More difficult is the identification of mixed model as we are going to see in the next application.

3.7.2 U.S. dollar interest rates

The second real application of identification and estimation techniques for multivariate time series models is to seven U.S. dollar interest rates with different terms to maturity, (6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years) from 30th November 1987 to 12th April 1990, which, excluding non trading days, gives 639 observations. This is the same dataset used by Tunnicliffe Wilson (1992) and obtained from Merryl Lynch. The relation between different assets whose only diversity is given by the different term

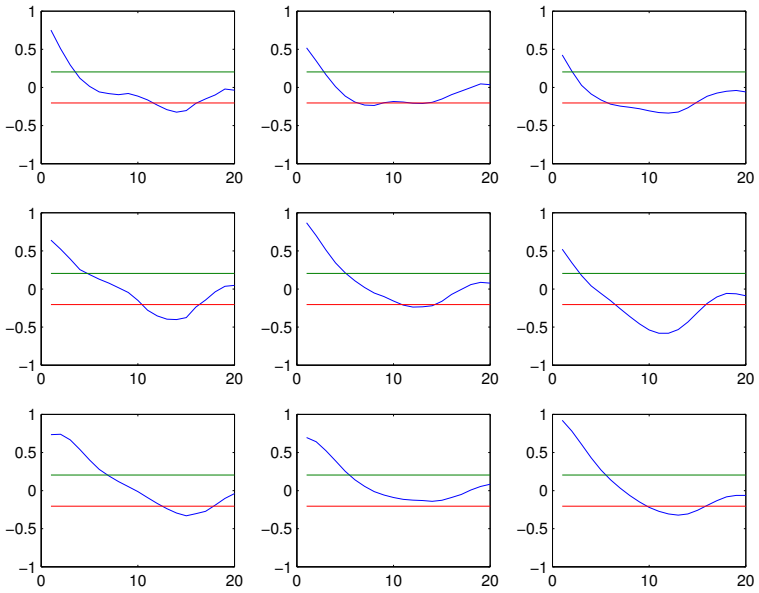


Figure 3.2: series correlations. This figure shows the correlations among the interest rates; on the abscissa there is the lag and on the ordinate the correspondent value of the correlation. In the three rows are showed the correlations of, respectively, (from above) REPO, treasury bills and bank loan interest rate with, respectively (columns) REPO treasury bills and bank loans interest rate so that on the main diagonal we have the autocorrelations of the three interest rates and cross-correlations elsewhere.

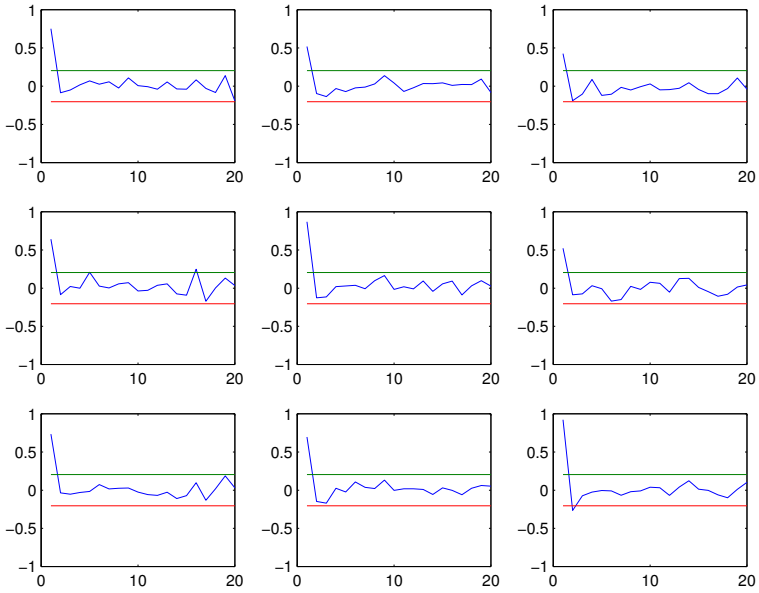


Figure 3.3: series partial correlations. This figure shows the partial correlations among the interest rates; on the abscissa there is the lag and on the ordinate the correspondent value of the partial correlation. In the three rows are shown the correlations of, respectively, (from above) REPO, treasury bills and bank loan interest rate with, respectively (columns) REPO treasury bills and bank loans interest rate so that on the main diagonal we have the partial autocorrelations of the three interest rates and cross-correlations elsewhere.

Table 3.1: Coefficient estimates and t-values for the VAR(2) model.

Dep. var.	Method	Indicator	$\phi_{a_{t-1}}$	$\phi_{a_{t-2}}$	$\phi_{b_{t-1}}$	$\phi_{b_{t-2}}$	$\phi_{c_{t-1}}$	$\phi_{c_{t-2}}$
a_t	LS	coeff.	0.6748	-0.0555	0.0784	0.0505	0.6457	-0.7104
		t-val.	5.4702	-0.4388	0.2082	0.1297	1.2080	-1.5734
	Y-W	coeff.	0.7411	-0.0893	-0.2112	0.2825	0.6968	-0.7827
		t-val.	5.6422	-0.6746	-0.5442	0.7438	1.4535	-1.7822
b_t	LS	coeff.	0.0509	-0.0249	1.2316	-0.2420	-0.1351	0.0317
		t-val.	1.2279	-0.5849	9.7219	-1.8489	-0.7515	0.2086
	Y-W	coeff.	0.0763	-0.0161	0.9966	-0.1149	-0.1246	-0.0051
		t-val.	1.3176	-0.2751	5.8211	-0.6858	-0.5894	-0.0262
c_t	LS	coeff.	0.0779	-0.0349	0.1874	-0.1613	1.3956	-0.5042
		t-val.	3.5529	-1.5557	2.8001	-2.3331	14.6938	-6.2847
	Y-W	coeff.	0.1088	0.0052	-0.0712	0.0549	1.1344	-0.3401
		t-val.	2.4341	0.1155	-0.5388	0.4251	6.9514	-2.2746

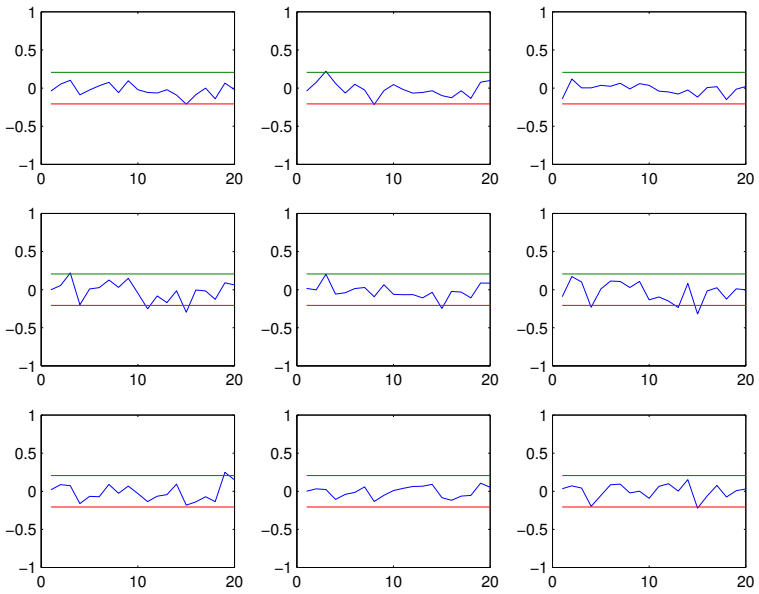


Figure 3.4: error correlations. The errors are obtained from the three regressions of the interest rates using ordinary least square estimation; on the abscissa there is the lag and on the ordinate the correspondent value of the correlation. In the three rows are shown the correlations of the errors obtained from the single regressions respectively of (from above): REPO, treasury bills and bank loan interest rate with, respectively (columns) REPO treasury bills and bank loans interest rate so that on the main diagonal we have the autocorrelations of the three interest rates and cross-correlations elsewhere.

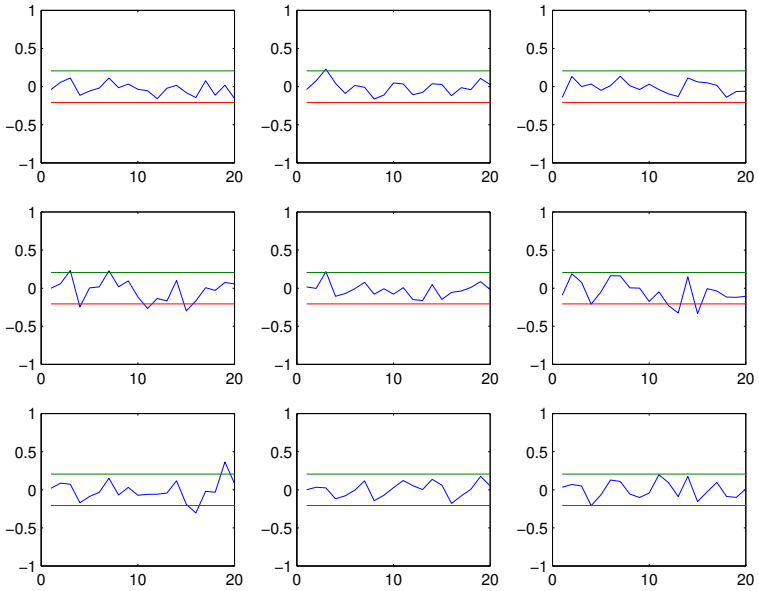


Figure 3.5: error partial correlations. The errors are obtained from the three regressions of the interest rates using ordinary least square estimation; on the abscissa there is the lag and on the ordinate the correspondent value of the correlation. In the three rows are showed the correlations of the errors obtained from the single regressions respectively of (from above): REPO, treasury bills and bank loan interest rate with, respectively (columns) REPO treasury bills and bank loans interest rate so that on the main diagonal we have the autocorrelations of the three interest rates and cross-correlations elsewhere.

to maturity, as in this case, is known as term structure. It plays an important role in economic policy and portfolio management strategies.

To understand this relation we build a multivariate ARMA model for the vector of actual forward values of U.S. dollar interest rate.

The series (fig. 3.6) are taken to be stationary. Over the time period of approximately three years it may be thought from the plots that they are integrated series, but knowledge of the longer term behaviour leads us to assume stationarity. The multivariate autocorrelations in figure 3.7 reflect this point: they are very slow to decay. Clearly a low order MA model is not appropriate.

The series partial autocorrelations in fig. 3.8 are all high at lag 1 (they are the lag 1 autocorrelations). There are significant values at lag 2, particularly in the partial cross-correlations of the first two rows. A pattern of values close to the lower limits is also evident at higher lags in these two rows. This is typical of patterns of autocorrelations and partial autocorrelations in the ARMA(1,1) model (see Box and Jenkins, 1976, p. 78, fig. 3.11).

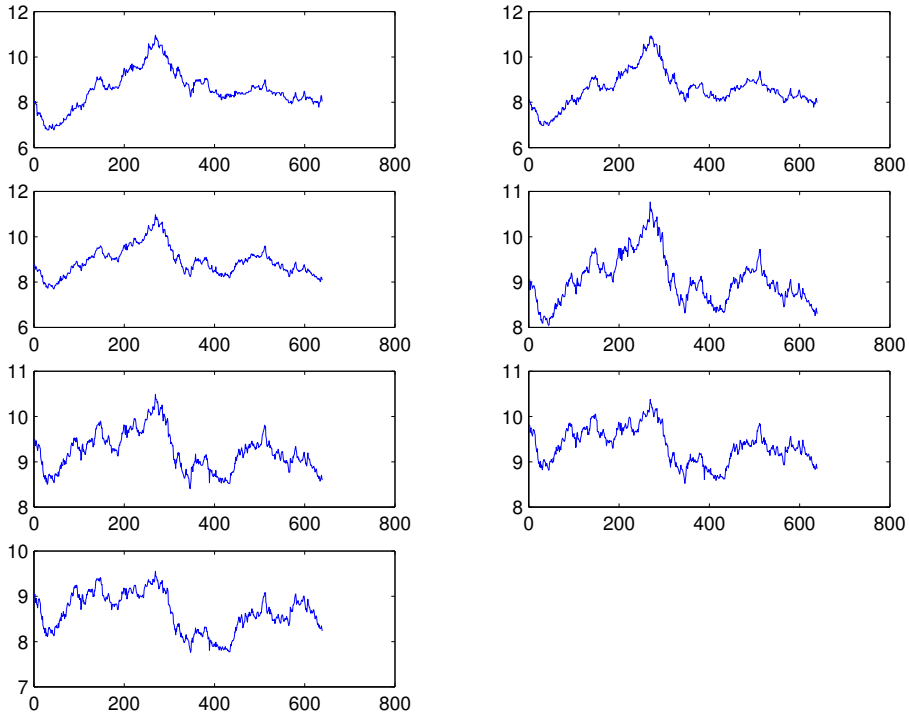


Figure 3.6: U.S. dollar interest rates. These are the time series of seven different terms to maturity of the U.S. dollar interest rate. On the abscissa there are the observations while on the ordinate the value of the interest rate. From above going from left to right are represented the following terms to maturities: 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years.

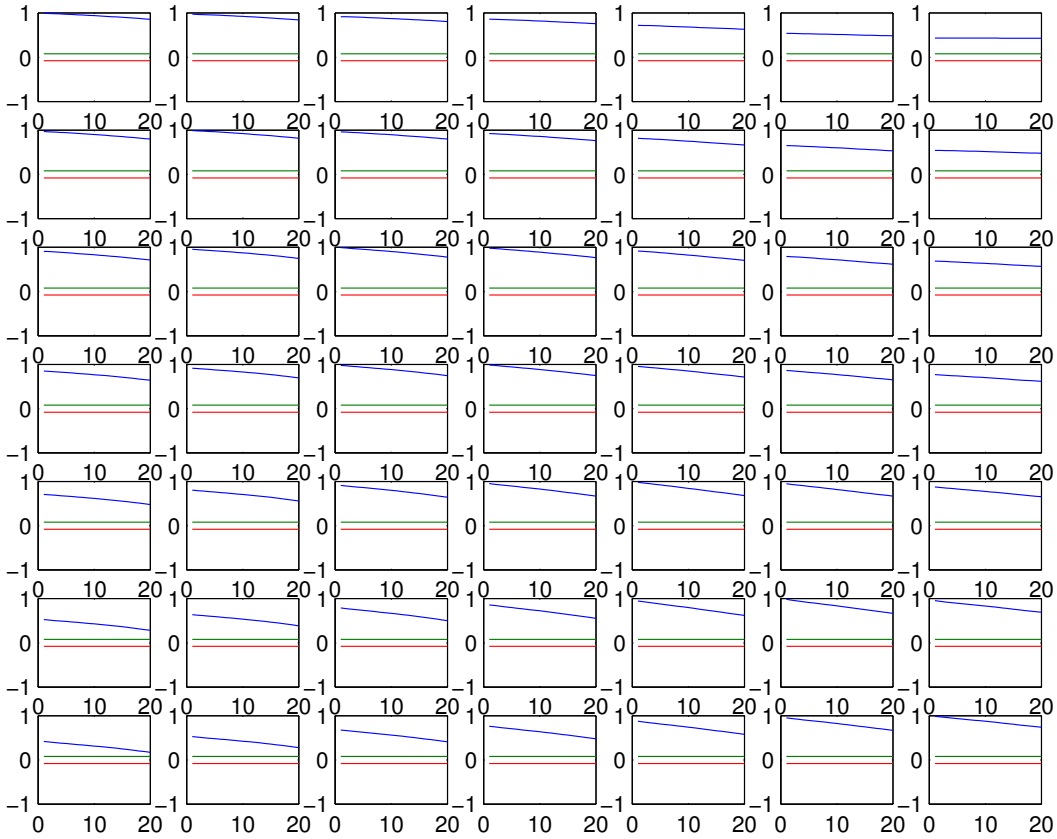


Figure 3.7: series correlations. This figure shows the correlations among the different terms to maturity of the U.S. dollar interest rate; on the abscissa there is the lag and on the ordinate the correspondent value of the correlation. In the seven rows are shown the correlations of, respectively, (from above) 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years with, respectively (columns) 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years so that on the main diagonal we have the autocorrelations and the cross-correlations elsewhere. The scale is chosen to show decay of the autocorrelation.

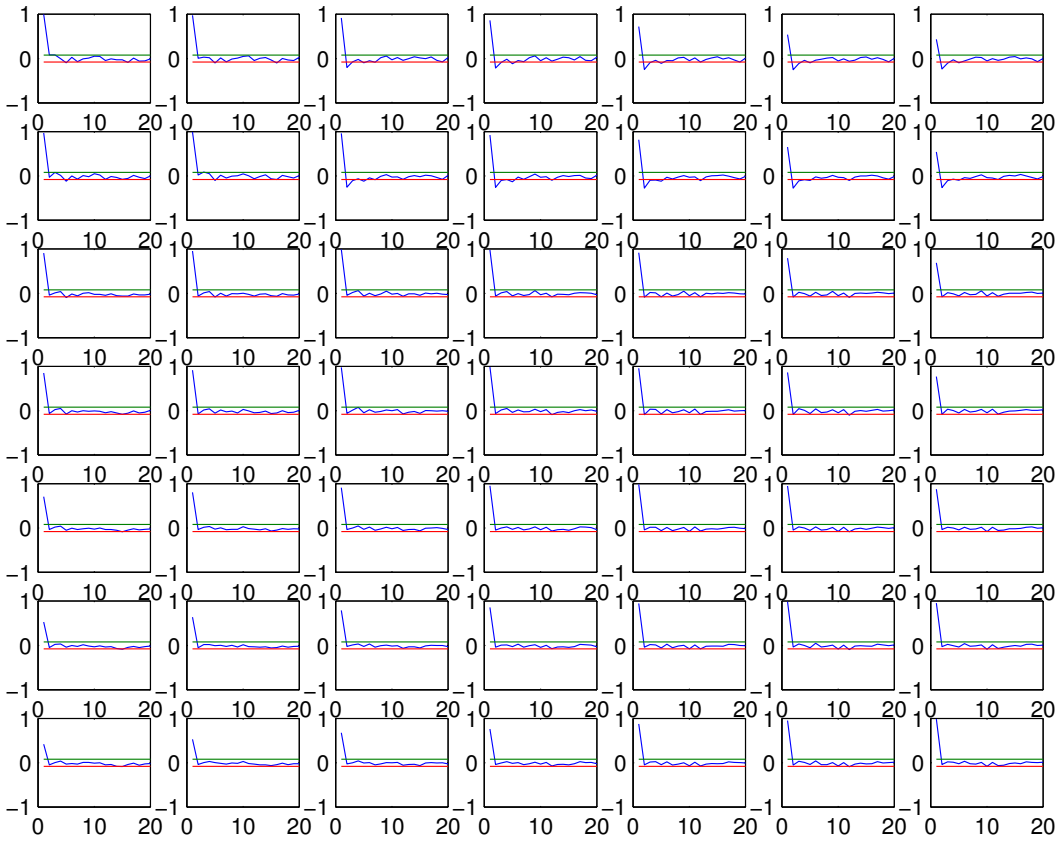


Figure 3.8: series partial correlations. This figure shows the partial correlations among the different terms to maturity of the U.S. dollar interest rate; on the abscissa there is the lag and on the ordinate the correspondent value of the partial correlation. In the seven rows are showed the partial correlations of, respectively, (from above) 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years with, respectively (columns) 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years so that on the main diagonal we have the partial autocorrelations and the partial cross-correlations elsewhere.

Using extended Yule-Walker equations we have identified a VARMA(1,1). We have estimated such a model using the method described in 3.6.3. We use $\mid \hat{\boldsymbol{V}} \mid$ to monitor convergence because the deviance (minus twice the log-likelihood) in 3.6.19 can be concentrated (minimised) with respect to $\boldsymbol{V}$ by taking

$$\hat{\boldsymbol{V}} = \frac{1}{n^*} \sum_{t=1}^{n} \boldsymbol{E}_t \boldsymbol{E}_t^T. \tag{3.7.1}$$

On substituting this into 3.6.19 we get

$$\begin{aligned} & n^* \log \mid \hat{\boldsymbol{V}} \mid \; + \; \operatorname{tr}\{\hat{\boldsymbol{V}}^{-1} n^* \hat{\boldsymbol{V}}\} \\ = \; & n^* \log \mid \hat{\boldsymbol{V}} \mid \; + \; \operatorname{tr}\{n\boldsymbol{I}\} \\ = \; & n^* \log \mid \hat{\boldsymbol{V}} \mid \; + \; nk. \end{aligned} \tag{3.7.2}$$

We have stopped the recursion after 30 iterations, having obtained convergence (see table 3.2).

Table 3.2: Convergence measure of estimation algorithm.

n. iterations	$\mid \hat{\boldsymbol{V}} \mid$
1	3.7874 E-18
2	4.0878 E-20
3	8.4462 E-21
10	1.9171 E-21
20	1.7931 E-21
30	1.7931 E-21

The resulting estimates for the autoregressive component are presented in table 3.3 while table 3.4 reports the estimates of the moving average component.

Table 3.3: Coefficients of the autoregressive matrix with * indicating t values greater than 1.96. For each equation (rows) are indicated the parameters of the independent variables (columns).

	column						
row	6m	1y	2y	3y	5y	7y	10y
6m	*0.90	0.01	*0.21	0.01	-0.15	-0.13	0.10
1y	*-0.05	*0.93	*0.24	-0.06	0.01	-0.24	*0.12
2y	*-0.06	0.00	* 1.23	-0.13	-0.08	-0.09	0.10
3y	*-0.07	0.00	*0.42	*0.61	-0.01	-0.11	*0.11
5y	*-0.09	0.01	*0.30	-0.16	*0.83	-0.05	0.08
7y	*-0.09	0.00	*0.35	-0.22	-0.02	*0.79	*0.13
10y	*-0.09	0.01	*0.26	-0.12	-0.06	-0.14	*1.08

Table 3.4: Coefficients of the moving average matrix. For each equation (rows) are indicated the parameters of the independent variables (columns).

	column						
row	6m	1y	2y	3y	5y	7y	10y
6m	0.34	0.08	-0.28	0.18	-0.55	-0.04	0.08
1y	-0.09	0.53	-0.42	0.18	-0.38	-0.22	0.13
2y	-0.03	-0.07	0.41	-0.13	-0.21	-0.27	0.15
3y	-0.05	-0.08	0.25	-0.02	-0.12	-0.24	0.07
5y	-0.04	-0.04	0.21	-0.27	0.25	-0.35	0.16
7y	-0.05	-0.06	0.33	-0.30	0.01	-0.13	0.13
10y	-0.08	-0.01	0.25	-0.18	0.03	-0.20	0.12

A confirmation of a proper identification of the model and a satisfactory estimation of the parameters is again confirmed by the correlations (fig. 3.9) and the partial correlations (fig. 3.10) of the residuals.

The residual series have high contemporaneous correlation, as shown in table 3.9. In later sections we shall re-examine this model and represent the contemporaneous correlation using structural modelling.

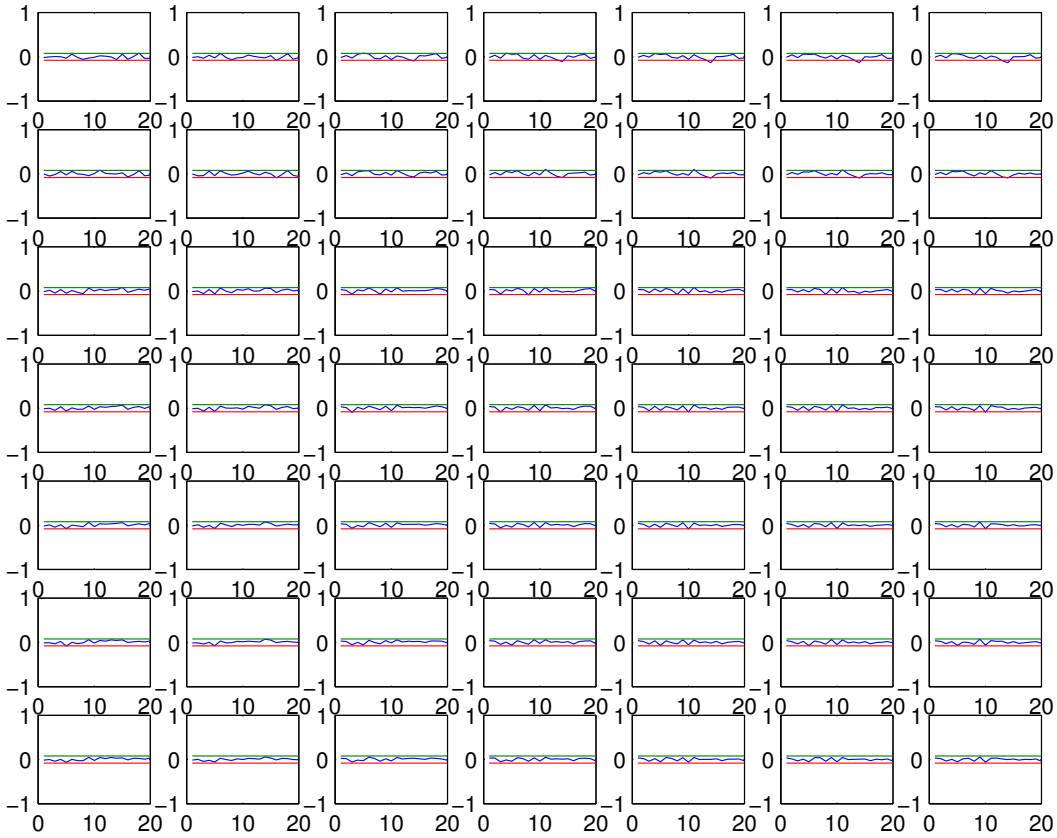


Figure 3.9: errors correlations. This figure shows the correlations among the estimation errors of different terms to maturity of the U.S. dollar interest rate; on the abscissa there is the lag and on the ordinate the correspondent value of the correlation. In the seven rows are showed the correlations of, respectively, (from above) 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years with, respectively (columns) 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years so that on the main diagonal we have the autocorrelations and the cross-correlations elsewhere.

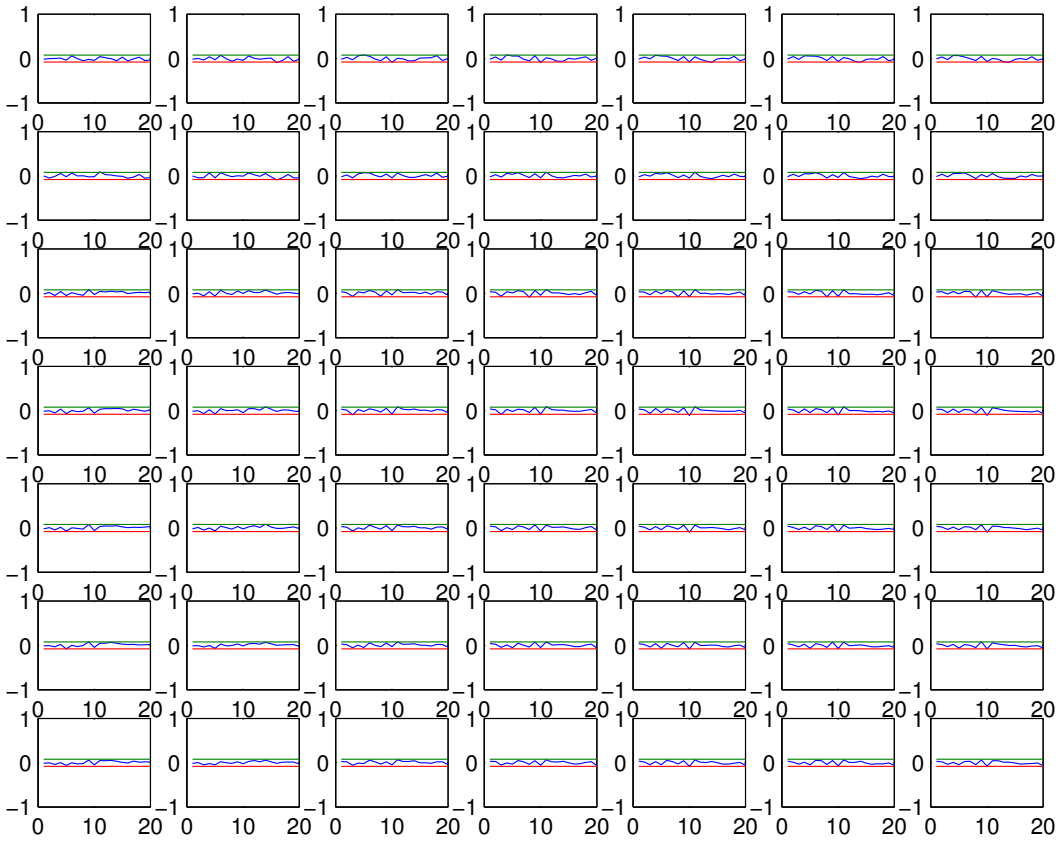


Figure 3.10: errors partial correlations. This figure shows the partial correlations among the estimation errors of different terms to maturity of the U.S. dollar interest rate; on the abscissa there is the lag and on the ordinate the correspondent value of the partial correlation. In the seven rows are showed the partial correlations of, respectively, (from above) 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years with, respectively (columns) 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years so that on the main diagonal we have the partial autocorrelations and the partial cross-correlations elsewhere.

Chapter 4

Graphical Modelling

In this chapter we shall introduce *Graphical Modelling*. It consists of methods and techniques using *Graph Theory* to model the relationships between random variables. It is natural to represent causal relationships among variables in a diagrammatic manner using *directed graphs* which use an arrow to link one variable to another which it affects causally.

In a statistical context, causality is difficult to establish, and is commonly a hypothesis used to interpret observed correlation between variables. To represent the relationships between variables characterised only by correlation, an undirected graph may be needed, the *conditional independence graph*. In the context of Gaussian random variables the key to conditional independence is the partial correlation among variables. A link between variables exists only if their partial correlation is non-zero. This chapter first defines conditional independence and the conditional independence graph and then introduces the concept of a *directed acyclic graph* and its relation with the conditional independence by graph *moralisation* in the Gaussian context. In the presentation of this topic we shall follow closely the texts by Diestel (1997) and Whittaker (1990).

We shall, in later chapters, use this graphical modelling approach as a first step in building structural models for multivariate time series. For this purpose it is important to understand which directed graphs (representing causality) are consistent with observed independence graphs (representing partial correlations).

In this chapter we do not consider the time dimension as an aspect of graphical modelling, but we illustrate the methods using the residuals from the interest rate model fitted in chapter 3.

4.1 Independence and conditional independence

In this section we present basic definitions of conditional independence. We also state the *block independence lemma*. This is the basis of the central result in graphical modelling, the *separation theorem*, presented in a later section.

4.1.1 Independence and conditional independence for events and random vectors

The conditional probability of an event A given an event B is

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \tag{4.1.1}$$

and is defined only if $\mathrm{P}(B) > 0$; then the events A and B are independent if

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B). \tag{4.1.2}$$

We use $A \perp\!\!\!\perp B$ to represent 4.1.2.

The central simplifying idea of graphical modelling is an extension of independence to conditional independence. We say that the events A and B are *conditionally independent* given the event C if and only if $\mathrm{P}(A \cap B \mid C) = \mathrm{P}(A \mid C)\mathrm{P}(B \mid C)$. We use the notation $A \perp\!\!\!\perp B \mid C$ to represent this.

The definition of independence and conditional independence of events can be extended to continuous random variables in terms of their marginal and joint density functions.

The conditional density of X given Y is

$$f_{X|Y}(x; y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

so X and Y are independent, written $X \perp\!\!\!\perp Y$, if

$$f_{XY}(x, y) = f_X(x) f_Y(y) \quad \forall x, y. \tag{4.1.3}$$

The definition of conditional independence of Y and Z given X, written $Y \perp\!\!\!\perp Z \mid X$ is

$$f_{YZ|X}(y, z; x) = f_{Y|X}(y; x) f_{Z|X}(z; x) \qquad \forall y, z, x \ s.t. \ f_X(x) > 0. \tag{4.1.4}$$

The block independence lemma which we now state, quoting Whittaker, is the basis of the separation lemma, presented later. Its importance is in taking pairwise conditional independence statements and constructing a group independence statement. It is used inductively in the separation lemma.

Proposition 4.1.1 *(block independence lemma). If (X, Y, Z_1, Z_2) is a partitioned random vector and $f(\cdot)$ is positive, then the following assertions are equivalent:*

$$Y \perp\!\!\!\perp (Z_1, Z_2) \mid X \tag{4.1.5}$$

$$Y \perp\!\!\!\perp Z_1 \mid (X, Z_2) \text{ and } Y \perp\!\!\!\perp Z_2 \mid (X, Z_1). \tag{4.1.6}$$

□

The second assertion follows almost directly from the first, using the definition we have given of conditional independence. It is the step from the second assertion to the first which is most important and requires careful consideration of the factorisation of $f_{XYZ_1Z_2}(x, y, z_1, z_2)$ implied by the second assertion.

4.2 Graphs

This section is required mainly to present terminology and notation such as "nodes" and "subgraphs" needed to describe graphical modelling. It follows introductory sections of Diestel quite closely, and figures 4.1 and 4.2 are taken from that text. The first of these is just an arbitrary graph, but constructed to illustrate the variety of connectivity, or lack of it, possible in a small graph. The second just illustrates the idea of subgraphs.

A graph is, formally, a pair $G = (V, E)$ where the elements of V are called *vertices* (or *nodes*) and the elements of E are called *edges* (or *lines*). To picture a graph we draw a circle (or a dot) for each vertex and join two of these circles by a line if the corresponding two vertices form an edge; how these circles and edges are placed is not relevant. An example of a graph is given in fig. 4.1. The number of vertices of a graph is its *order*. We say that a vertex v is *incident* with an edge e and conversely e is an edge *at* v; two vertices incident with an edge are its *endvertices* or *ends*, and an edge *joins* its ends.

Two vertices x, y of G are *adjacent*, or *neighbours*, if xy is an edge of G; this is the case for vertices 2 and 5 in the graph in figure 4.1. The *degree* or *valency* $d_G(v)$, or simply $d(v)$ of a vertex v is the number of edges at v; by our previous definition of a graph, this is equal to the number of neighbours of v. A vertex of degree 0 is *isolated*. Two edges $e \neq f$ are *adjacent* if they have an end in common, like edges 2-5 and 2-7 which have the vertex 2 in common in figure 4.1. If all the vertices of G are pairwise adjacent, then G is *complete*. A complete graph on n vertices is a K^n; a K^3 is called *triangle*.

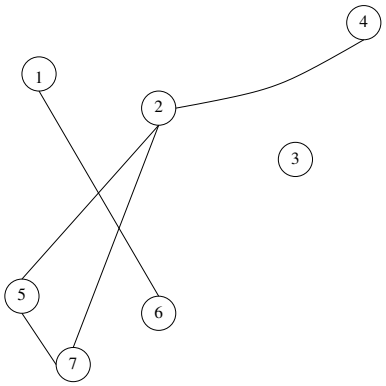


Figure 4.1: Graph on V={1,...,7} with edge set E={{1,6},{2,4},{2,5},{2,7},{5,7}}.

Let $G \cup G' = (V \cup V', E \cup E')$ and $G \cap G' = (V \cap V', E \cap E')$. If $G \cap G' = 0$ then G and G' are *disjoint*. If $V' \subseteq V$ and $E' \subseteq E$, the G' is a *subgraph* of G (and G a *supergraph* of G'), written as $G' \subseteq G$. In this case if G' contains all the edges $xy \in E$ with $x, y \in V'$, then G' is an induced subgraph of G; figure 4.2 can clarify this concept.

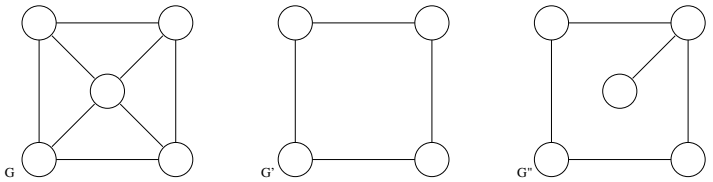


Figure 4.2: G' and G'' are both subgraphs of G; G' is an induced subgraph of G while G'' is not.

A path is a non-empty graph $P = (V, E)$ of the form

$$V = \{x_0, x_1, \ldots, x_k\} \qquad E = \{x_0x_1, x_1x_2, \ldots, x_{k-1}x_{k-2}\}$$

where the x_i are all distinct. The vertices x_0 and x_k are linked by P and are called its ends; the vertices $x_1, \ldots, x_{k-1}$ are the *inner* vertices of P. The number of edges of a path is its *length*, and the path of length k is denoted by P^k (see figure 4.3), where k can be zero; in this case $P^0 = K^1$.

If $P = x_0, \ldots, x_{k-1}$ is a path and $k \geq 3$, then the graph $C = P + x_{k-1}x_0$ is called a *cycle*. For example the graph in figure 4.4 is a cycle obtained from the path in figure 4.3.

A cycle is characterised by its *length* which is the number of its edges (e.g. the cycle in fig. 4.4 has length 4). If an edge joins two vertices of a cycle but it is not part of it, it is called a *chord*. A non-empty graph G is called *connected* if any two of its vertices are linked by a path in G; if we consider, for example, the graph in figure 4.1, it is not

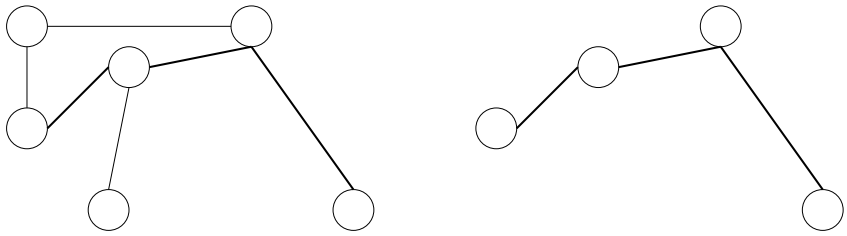


Figure 4.3: Path. The graph on the right is a path ($P^4(E', V')$) of the graph ($G(E, V)$ on the left.

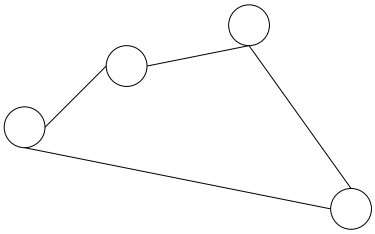


Figure 4.4: Cycle obtained from the path in figure 4.3

connected because of vertex 3.

A *directed graph* or *digraph* is a pair (V, E) of disjoint sets of vertices and edges together with two maps, init:$E \to V$ and ter:$E \to V$, assigning to every edge e an *initial vertex*, init(e), and a *terminal vertex*, ter(e). The edge is said to be *directed* from init(e) to ter(e). Initial vertices are also indicated as *parents* of the terminal vertices, pa(ter(e)).

We will use undirected graphs such as these to represent conditional independence relationships between variables. To represent causal relationships we need directed graphs.

A directed graph may have several edges between the same two vertices x, y. Such edges are called *multiple edges* and if they have the same direction, they are *parallel*. if init(e)=ter(e), the edge e is called a *loop*. A directed graph D is an orientation of an *undirected graph* G if $V(D) = V(G)$ and $E(D) = E(G)$ and if {init(e),ter(e)}=$\{x, y\}$ for every edge $e = xy$. Intuitively, such an *oriented graph* arises from an undirected graph simply by directing every edge from one of its ends to the other. Put differently, oriented graphs are directed graphs without loops and multiple edges.

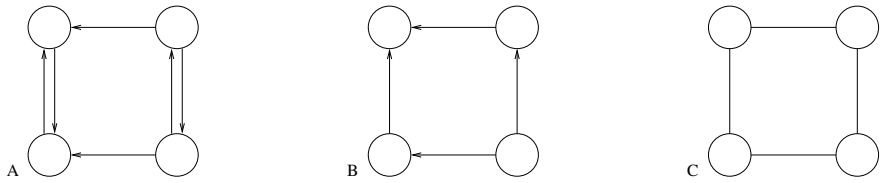


Figure 4.5: A: directed graph; B: oriented graph; C: undirected graph

In this thesis we shall not consider multiple edges and loops and by directed graphs we shall mean oriented graphs, consistently with the graphical modelling terminology.

4.3 Conditional independence graphs

If we have a set of variables $Y_i, ..., Y_n$ we can represent the relation among them by a graph where each variable is represented by a vertex and where two vertices are linked by an edge if and only if the corresponding variables are conditionally dependent given the remaining variables; such a graph is called a *conditional independence graph* (CIG). To make things clearer let us consider the conditional independence graph in figure 4.6:

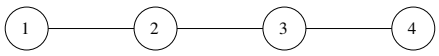


Figure 4.6: conditional independence graph

variables 3 and 4 are conditionally dependent given 1 and 2 while variables 1 and 3 are not because they are neighbours. This example would probably make you wonder if we really need to consider variable 4 in the conditioning set to state the conditional independence of variables 1 and 3; that is, in a nutshell, can we say $1 \perp\!\!\!\perp 3 \mid 2$ instead of $1 \perp\!\!\!\perp 3 \mid \{2, 4\}$? Because of the separation theorem it is possible to give a positive answer to such a question. The separation theorem is the consequence of successive applications of the block independence lemma (proposition 4.1.1).

4.3.1 Separation theorem

In this section we state the theorem as presented by Whittaker (1990, pp. 64–67) and summarise its proof which is given by Whittaker. Formally, the content of the separation theorem is given by

Theorem 4.3.1 *(separation theorem). If X_a, X_b and X_c are vectors containing disjoint subsets of variables from X, and if, in the independence graph of X, each vertex in X_a is separated from each vertex in X_c by the subset X_b, then*

$$X_a \perp\!\!\!\perp X_c \mid X_b$$

□

We are given in the statement of this theorem conditions which may be expressed simply in block terms by the CIG in figure 4.7

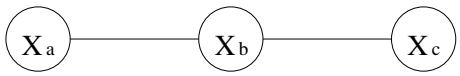


Figure 4.7: CIG.

where, for example, the link between X_a and X_b represents all the edges between variables in X_a and X_b. The graph is based only on pairwise conditional independence statements where the set of conditioning variables for each pair is the whole set of variable excluding that pair.

The aim of the theorem is to show that for the above diagram conditional independence applies between any subset of X_a and X_c where now the set of conditioning variables is only those of X_b. The key step is the application of the block independence lemma. This is applied successively to remove variables in the conditioning sets in X_a and X_c until the only remaining conditioning variables are in X_b. The important point is that this can be done without forming any new links between X_a and X_c as the conditioning set changes. The proof depends upon careful choice of the variables selected for applying the block independence lemma.

4.4 Directed acyclic graphs

We can give a direction to the edges of a CIG introducing, in this way, the notion of causality in the relationships between variables.

Doing so we have to face the troublesome interpretation of cycles.

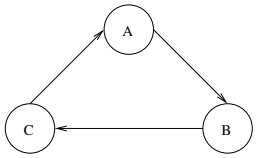


Figure 4.8: Directed cyclic graph.

The joint density function implied by the graph in figure 4.8 is

$$f_{ABC}(\cdot) = f_{B|A}(\cdot) f_{C|B}(\cdot) f_{A|C}(\cdot)$$

but apart from very special cases such a density function is not well defined hence we shall not consider the possibility of the existence of cycles. This is equivalent to a complete ordering of the vertices in the graph itself (see Whittaker, 1990, p. 72).

The main difference between directed and undirected independence graphs is that for an undirected graph, the independence statements are statements about a single joint distribution while for a directed graph they are statements about a sequence of marginal distributions. Though the two are linked as from the marginal distributions we may derive the conditional ones and then the joint distributions by the *recursive factorisation identity*

$$f_{1,2,\ldots,k}(\cdot) = f_{k|S(k)\setminus\{k\}}(\cdot) f_{\{k-1\}|S(k-1)\setminus\{k-1\}}(\cdot) \ldots f_{2|1}(\cdot) f_1(\cdot) \tag{4.4.1}$$

4.4.1 Wermuth condition

We indicate a directed graph as $G^<(V, E^<)$. Intuitively we may think that its equivalent undirected graph $G^U(V, E^U)$, in terms of conditional independence interpretation, would be obtained by just substituting the directed edges with undirected ones but it is not so simple. As the notation may suggest the vertices set, V, is the same but some changes may intervene in the edge set. We use an example to explain what happens when we want to obtain $G^U(V, E^U)$.

Example 4.4.1 . *Consider the directed graph D_1 in figure 4.9 It is defined by the*

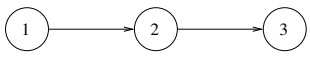


Figure 4.9: graph D_1

factorisation $f_{1,2,3}(\cdot) = f_{3|2}(\cdot) f_{2|1}(\cdot) f_1(\cdot)$.

Now consider the graph where the directed edges are substituted by undirected ones (figure 4.10), we call it U_1.

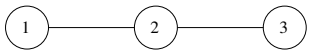


Figure 4.10: graph U_1

It is defined by the independence $3 \perp\!\!\!\perp 1 \mid 2$, i.e.

$$f_{1,3|2}(\cdot) = f_{3|2}(\cdot) f_{1|2}(\cdot).$$

To prove the equivalence of D_1 *and* U_1 *we have to verify that* $D_1 \Longrightarrow U_1$ *and that* $U_1 \Longrightarrow D_1$*. For this example*

$$f_{3,1|2}(\cdot) = f_{3|2}(\cdot)f_{1|2}(\cdot) \Longrightarrow f_{3,1,2}(\cdot) = f_{3|2}(\cdot)f_{1|2}(\cdot)f_2(\cdot) = f_{3|2}(\cdot)f_{1,2}(\cdot).$$

Also

$$f_{3|2}(\cdot)f_{2|1}(\cdot)f_1(\cdot) = f_{3|2}(\cdot)f_{2,1}(\cdot)$$

so these are equivalent. □

In the next example we shall consider a similar but less trivial situation.

Example 4.4.2 . *Consider the graph* D_2

Figure 4.11: graph D_2

It is defined by the independences: d_1) $4 \perp\!\!\!\perp 2 \mid \{1,3\}$, d_2) $4 \perp\!\!\!\perp 1 \mid \{2,3\}$ *and* d_3) $3 \perp\!\!\!\perp 1 \mid 2$*. Substituting the arrows with straight lines we obtain the graph* U_2 *Such a graph is*

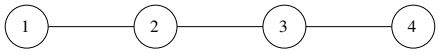


Figure 4.12: graph U_2

described by the conditional independences: u_1) $1 \perp\!\!\!\perp 4 \mid \{2,3\}$*;* u_2) $2 \perp\!\!\!\perp 4 \mid \{1,3\}$, u_3) $3 \perp\!\!\!\perp 1 \mid \{2,4\}$.

Let us first prove that $D_2 \Longrightarrow U_2$*. Using the recursive factorisation identity the density function of* D_2 *can be written as*

$$\begin{aligned} f_{1,2,3,4}(\cdot) &= f_{4|3}(\cdot)f_{3|2}(\cdot)f_{2|1}(\cdot)f_1(\cdot) && (4.4.2)\\ &= g_4(\cdot)g_1(\cdot) && (4.4.3)\end{aligned}$$

where $g_4(\cdot) = f_{4|3}(\cdot)f_{3|2}(\cdot)$ *and* $g_1(\cdot) = f_{2|1}(\cdot)f_1(\cdot)$ *which, given 3 and 2, are respectively functions of 4 and 1 and hence because of the factorisation criterion* $4 \perp\!\!\!\perp 1 \mid \{3,2\}$*. Similarly we can factorise 4.4.2 as*

$$f_{1,2,3,4}(\cdot) = h_4(\cdot)h_2(\cdot) \tag{4.4.4}$$

where $h_4(\cdot) = f_{4|3}(\cdot)$ *and* $h_2(\cdot) = f_{3|2}(\cdot)f_{2|1}(\cdot)f_1(\cdot)$ *then*

$$4 \perp\!\!\!\perp 2 \mid \{1,3\}. \tag{4.4.5}$$

Also

$$f_{1,2,3,4}(\cdot) = m_3(\cdot)m_1(\cdot) \tag{4.4.6}$$

where $m_3(\cdot) = f_{4|3}(\cdot)f_{3|2}(\cdot)$ *and* $m_1(\cdot) = f_{2|1}(\cdot)f_1(\cdot)$. *This factorisation is very similar to the first one, with* $h_4(\cdot)$ *and* $h_2(\cdot)$, *the difference is that we consider 2 and 4 given, hence*

$$3 \perp\!\!\!\perp 1 \mid 2 \tag{4.4.7}$$

proving in this way that D_2 *implies* U_2.

Now we have to prove that $U_2 \Longrightarrow D_2$.

The independence statement d_3 *gives a useful information, in fact*

$$\begin{aligned} 3 \perp\!\!\!\perp 1 \mid 2 \Longrightarrow f_{1,2,3}(\cdot) &= f_{3|\{2,1\}}(\cdot)f_{2,1}(\cdot) \\ &= f_{3|2}(\cdot)f_{2,1}(\cdot). \end{aligned} \tag{4.4.8}$$

Because of the block independence lemma, independence statements d_1 *and* d_2 *are equivalent to* $4 \perp\!\!\!\perp \{1,2\} \mid 3$, *then*

$$\begin{aligned} 4 \perp\!\!\!\perp \{1,2\} \mid 3 \Longrightarrow f_{4,3,2,1}(\cdot) &= f_{4|\{3,2,1\}}(\cdot)f_{3,2,1}(\cdot) \\ &= f_{4|3}(\cdot)f_{3,2,1}(\cdot) \end{aligned} \tag{4.4.9}$$

then substituting 4.4.8 in 4.4.9 we have

$$f_{4,3,2,1}(\cdot) = f_{4|3}(\cdot)f_{3|2}(\cdot)f_{2,1}(\cdot) \tag{4.4.10}$$

and because always

$$f_{2,1}(\cdot) = f_{2|1}(\cdot)f_1(\cdot) \tag{4.4.11}$$

finally we have

$$f_{4,3,2,1}(\cdot) = f_{4|3}(\cdot)f_{3|2}(\cdot)f_{2|1}(\cdot)f_1(\cdot) \tag{4.4.12}$$

proving in this way that U_2 *implies* D_2. □

In the last two examples we have obtained the equivalent undirected graph simply by taking out the direction from the edges. But it doesn't always work in this way. Let us consider another example

Example 4.4.3 . *The density function of graph* D_3 *is*

$$f_{1,2,3}(\cdot) = f_{3|\{1,2\}}(\cdot)f_{1,2}(\cdot). \tag{4.4.13}$$

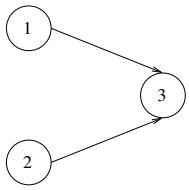


Figure 4.13: graph D_3

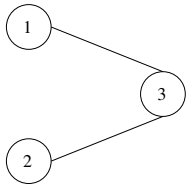


Figure 4.14: graph U_3

Substituting arrows with lines in D_3, we obtain U_3 which is described by the independence statement

$$1 \perp\!\!\!\perp 2 \mid 3. \tag{4.4.14}$$

In this way marginal independence would imply conditional independence, but this is not correct (see Whittaker, pp. 24–30). □

In general, every time we have a directed graph with a subgraph like D_3 we cannot obtain the equivalent undirected graph simply substituting the arrows with lines. When no subgraph has the configuration of D_3 the directed graph is said to satisfy the *Wermuth condition.*

To obtain the equivalent undirected graph when the Wermuth condition is not satisfied we have to eliminate all the forbidden configurations like D_3 by adding edges to link parents.

If we consider again the graph D_3 in the example 4.4.3, linking parents would lead to graph M_3 in figure 4.15

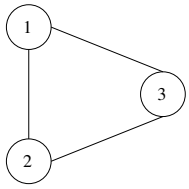


Figure 4.15: graph M_3

Lauritzen and Spiegelhalter (1988) introduced the verb marrying instead of linking and defined graphs like M_3, where parents are married, moral. Such a graph is equivalent, from a conditional independence point of view, to its associated directed graph, in fact equation 4.4.1, because of the separation theorem, can be expressed as

$$f_{1,2,\ldots,k}(\cdot) = f_{k|\{\mathrm{pa}(k)\}}(\cdot) f_{\{k-1\}|\{\mathrm{pa}(k-1)\}}(\cdot) \ldots f_1(\cdot) \tag{4.4.15}$$

and by choosing an appropriate function $g(\cdot)$ we can rewrite it as

$$f_{1,2,\ldots,k}(\cdot) = g_{k\cup\{\mathrm{pa}(k)\}}(\cdot) g_{\{k-1\}\cup\{\mathrm{pa}(k-1)\}}(\cdot) \ldots g_1(\cdot). \tag{4.4.16}$$

This joint density function describes a graph whose edge set is the same of the associated moral graph, that is: every vertex is connected to its parents and parents are connected. Conversely applying the recursive factorisation identity to the expansion 4.4.16 we obtain all the pairwise conditional independences.

4.4.2 Demoralisation of CIG's

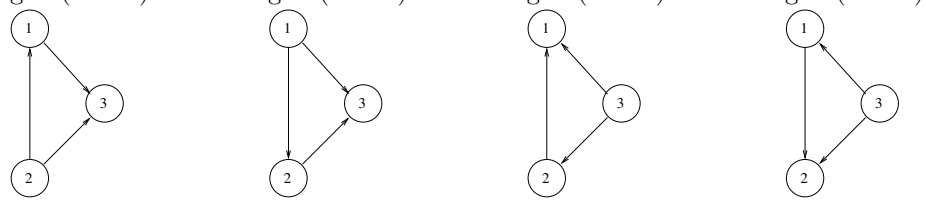
In the previous section we have seen how it is possible to pass from a DAG to a CIG using moralisation. Nevertheless in many situations we are interested in going from a CIG to a DAG. In other words once we know the relations among the variables involved in a system as described by a CIG, we may be interested in characterising the causal structure as described by a DAG. For this purpose we have to consider what possible DAG's may give rise to the observed CIG. To construct such a possible DAG we assign directions to the edges and take out the edges which might have arisen by moralisation.

This process has been considered by some authors, but there is generally no unique resulting DAG (Spirtes *et al.*, 1993 and Pearl, 1988), in contrast to the unique CIG of moralization. For the identification of our structural models the process is central. We have therefore developed this area by considering the model criteria, such as likelihoods, which are needed to select the DAG representations. Consistently with the terminology of Lauritzen and Spiegelhalter we have given the name "demoralisation" to the act of taking out moral edges and assigning directions.

As an example, if we demoralise $M3$, giving a direction to the edges and then taking out the edge linking pa(3), i.e. vertices 1 and 2, we would get D_3.

Example 4.4.4 . *Suppose we have a CIG like $M3$ (fig. 4.15). The number of possible directed graphs with the same number of edges is computed considering that every edge*

Figure(4.16.1) Figure(4.16.2) Figure(4.16.3) Figure(4.16.4)



Figure(4.16.5) Figure(4.16.6) Figure(4.16.7) Figure(4.16.8)

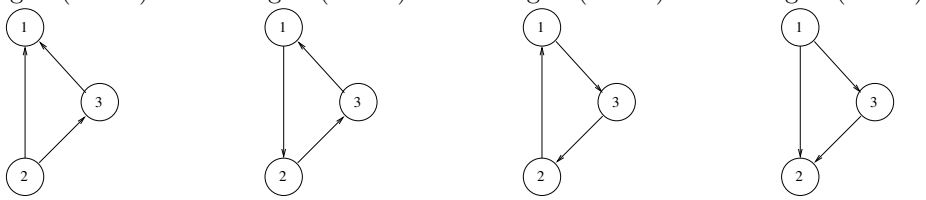


Figure 4.16: Possible equivalent directed graph for $M3$ with the same number of edges.

can assume two different directions; hence it is 2^n *where* n *is the number of edges. In our case it is* $2^3 = 8$ *(see figure 4.16). Two of these graphs are cyclical (figures 4.16.6 and 4.16.7) and hence will be excluded. So we have now six possible models and we have to decide which of them is the most adequate. There are some strategies, that will be described later, to take such a decision. Once we have made our choice and selected a model we have to demoralise it taking out edges whose presence is due to moralisation. For example, if we would have selected the model in figure 4.16.2, it is possible that the edge linking vertices 1 and 2 is a moral one; if this is the case, and again there are methods to assess that we shall take it out demoralising in this way the graph.* □

Demoralisation will play a central role in the rest of this thesis and we shall give a direction to the edges comparing alternative models with likelihood based methods which will be described in the next chapter.

There are cases, where there is not any suitable DAG equivalent to a given CIG (see Pearl, 1988, pp. 130–131). In such a case the situation can be recovered by introducing extra variables.

4.5 Gaussian CIG models and the inverse variance lemma

Theory, so far, has been developed in the context of general distributions but now we wish to make the common simplifying assumption that our data is Gaussian or distributed

according a multivariate normal distribution i.e.

$$\boldsymbol{X} \sim MVN(\mu, \boldsymbol{V}).$$

In that case the existence of conditional independence, which is needed to define graphical models, is determined by examining the partial correlations between pair of variables, i.e.

$$x_i \perp\!\!\!\perp x_j \iff \tau_{i,j} = \text{Corr}\,(x_i, x_j \mid X_c) = 0$$

where $X_c = \boldsymbol{X} \backslash \{x_i, x_j\}$ and $\text{Corr}\,(x_i, x_j)$ is known as the partial correlation between x_i and x_j.

Knowing $\boldsymbol{V}$, the conditional independence graph of $\boldsymbol{X}$ can be easily constructed by calculating all the necessary partial correlations using the following lemma.

Theorem 4.5.1 *(inverse variance lemma). Consider the variables vector $\boldsymbol{X}$ where $\boldsymbol{X} \sim$ MVN($\boldsymbol{\mu}$,$\boldsymbol{V}$). Let $\boldsymbol{W} = \boldsymbol{V}^{-1}$ and $\tau_{ij} = -\frac{w_{ij}}{\sqrt{w_{ii}, w_{jj}}}$ where $w_{ii}w_{jj}$ and w_{ij} are elements of $\boldsymbol{W}$; then $\tau_{i,j} = corr(x_i, x_j \mid \boldsymbol{X} \backslash \{x_i, x_j\})$*

Remark: See also Mardia, Kent and Bibby (1980) chapter 6 for definitions of the sample partial correlation and appendix A2.4 for statement of partitioned matrix inverses.
Proof: Define $\boldsymbol{X}_c = \boldsymbol{X} \backslash \{x_i, x_j\}$ and $\hat{x}_i(\boldsymbol{X}_c)$ and $\hat{x}_j(\boldsymbol{X}_c)$ as the conditional expectations of x_i and x_j given $\boldsymbol{X}_c$, that implies

$$\begin{aligned} x_i &= \hat{x}_i(\boldsymbol{X}_c) + e_i \\ x_j &= \hat{x}_j(\boldsymbol{X}_c) + e_j \end{aligned} \tag{4.5.1}$$

then

$$\text{corr}(x_i, x_j \mid \boldsymbol{X}_c) = \text{corr}(e_i, e_j). \tag{4.5.2}$$

Without loss of generality, take x_1 and x_2 and let $\boldsymbol{Y} = (x_1, x_2)^T$ and $\boldsymbol{Z} = \boldsymbol{X}_c^T$, then the covariance matrix of all the variables, $\boldsymbol{V}$, can be partitioned as

$$\text{Var}\,(\boldsymbol{X}) = \boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_{YY} & \boldsymbol{V}_{YZ} \\ \boldsymbol{V}_{ZY} & \boldsymbol{V}_{ZZ} \end{bmatrix} \tag{4.5.3}$$

where $\boldsymbol{V}_{YZ}$ is the matrix of the correlations between the elements of $\boldsymbol{Y}$ and the elements of $\boldsymbol{Z}$ and so on.

From 4.5.1 we have

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{E} \qquad \boldsymbol{E} \perp\!\!\!\perp \boldsymbol{Z} \tag{4.5.4}$$

where $\boldsymbol{E}$ is the vector of errors $e_i, e_j, \ldots$, then

$$\text{Cov}\,(\boldsymbol{Z}, \boldsymbol{Y}) = \text{Cov}\,(\boldsymbol{Z}, \boldsymbol{AZ} + \boldsymbol{E}) = \boldsymbol{A}\text{Cov}\,(\boldsymbol{Z}, \boldsymbol{Z}) \tag{4.5.5}$$

and considering 4.5.3

$$\boldsymbol{V}_{ZY} = \boldsymbol{AV}_{ZZ} \Longleftrightarrow \boldsymbol{A} = \boldsymbol{V}_{ZY}\boldsymbol{V}_{ZZ}^{-1}. \tag{4.5.6}$$

Also,

$$\begin{aligned} \text{Var}\,(\boldsymbol{E}) &= \text{Cov}\,(\boldsymbol{E}, \boldsymbol{E}) \\ &= \text{Cov}\,(\boldsymbol{Y} - \boldsymbol{AZ}, \boldsymbol{E}) \\ &= \text{Cov}\,(\boldsymbol{Y}, \boldsymbol{E}) \\ &= \text{Cov}\,(\boldsymbol{Y}, \boldsymbol{Y} - \boldsymbol{AZ}) \\ &= \boldsymbol{V}_{YY} - \boldsymbol{V}_{ZY}\boldsymbol{A}^T. \end{aligned} \tag{4.5.7}$$

Then, substituting A as in 4.5.6 we have

$$\text{Var}\,(\boldsymbol{E}) = \boldsymbol{V}_{YY} - \boldsymbol{V}_{ZY}\boldsymbol{V}_{ZZ}^{-1}\boldsymbol{V}_{YZ}. \tag{4.5.8}$$

Let

$$\text{Var}\,(x_1, x_2 \mid \boldsymbol{Z}) = \text{Var}\,(\boldsymbol{E}) = \boldsymbol{V}_{YY} - \boldsymbol{V}_{ZY}\boldsymbol{V}_{ZZ}^{-1}\boldsymbol{V}_{YZ} = \boldsymbol{U} \tag{4.5.9}$$

then

$$\text{Corr}\,(x_1, x_2 \mid \boldsymbol{Z}) \equiv \frac{u_{12}}{\sqrt{u_{11}u_{22}}}. \tag{4.5.10}$$

Now let

$$\boldsymbol{W} = \boldsymbol{V}^{-1} = \begin{bmatrix} \boldsymbol{W}_{YY} & \boldsymbol{W}_{YZ} \\ \boldsymbol{W}_{ZY} & \boldsymbol{W}_{ZZ} \end{bmatrix} \tag{4.5.11}$$

then what we need to prove is

$$\boldsymbol{W}_{YY} = \boldsymbol{U}^{-1}. \tag{4.5.12}$$

To do that let us start considering

$$\boldsymbol{VW} = \boldsymbol{I}_n \tag{4.5.13}$$

where $\boldsymbol{I}_n$ is a $n \times n$ identity matrix. We can rewrite 4.5.13 as

$$\begin{bmatrix} \boldsymbol{V}_{YY} & \boldsymbol{V}_{YZ} \\ \boldsymbol{V}_{ZY} & \boldsymbol{V}_{ZZ} \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_{YY} & \boldsymbol{W}_{YZ} \\ \boldsymbol{W}_{ZY} & \boldsymbol{W}_{ZZ} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_2 & 0 \\ 0 & \boldsymbol{I}_{n-2} \end{bmatrix}. \tag{4.5.14}$$

From 4.5.14

$$\boldsymbol{V}_{YY}\boldsymbol{W}_{YY} + \boldsymbol{V}_{YZ}\boldsymbol{W}_{ZY} = \boldsymbol{I}_2 \tag{4.5.15}$$

$$\boldsymbol{V}_{ZY}\boldsymbol{W}_{YY} + \boldsymbol{V}_{ZZ}\boldsymbol{W}_{ZY} = 0. \tag{4.5.16}$$

Rearranging 4.5.16

$$\boldsymbol{W}_{ZY} = -\boldsymbol{V}_{ZZ}^{-1}\boldsymbol{V}_{ZY}\boldsymbol{W}_{YY} \tag{4.5.17}$$

Then, considering 4.5.15 and 4.5.17

$$\boldsymbol{V}_{YY}\boldsymbol{W}_{YY} - \boldsymbol{V}_{YZ}\boldsymbol{V}_{ZZ}^{-1}\boldsymbol{V}_{ZY}\boldsymbol{W}_{YY} = \boldsymbol{I} \tag{4.5.18}$$

that is

$$(\boldsymbol{V}_{YY} - \boldsymbol{V}_{YZ}\boldsymbol{V}_{ZZ}^{-1}\boldsymbol{V}_{ZY})\boldsymbol{W}_{YY} = \boldsymbol{I} \tag{4.5.19}$$

and then considering 4.5.9 and 4.5.19

$$\boldsymbol{U}\boldsymbol{W}_{YY} = \boldsymbol{I} \iff \boldsymbol{U} = \boldsymbol{W}_{YY}^{-1}. \tag{4.5.20}$$

So now we have just to prove that

$$\frac{u_{12}}{\sqrt{u_{11}u_{22}}} = \frac{-w_{12}}{\sqrt{w_{11}w_{22}}}. \tag{4.5.21}$$

Let

$$\boldsymbol{U} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \text{ so that } \frac{u_{12}}{\sqrt{u_{11}u_{22}}} = \rho \tag{4.5.22}$$

then

$$\boldsymbol{W} = \boldsymbol{U}^{-1} = \frac{1}{\lambda}\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} \qquad \lambda = \sigma_1^2\sigma_2^2(1-\rho^2) \tag{4.5.23}$$

hence

$$w_{11} = \frac{\sigma_1^2}{\lambda}; \quad w_{22} = \frac{\sigma_2^2}{\lambda}; \quad \text{and} \quad w_{12} = \frac{-\rho\sigma_1\sigma_2}{\lambda}.$$

Finally substituting these values into the RHS of 4.5.21 and considering 4.5.22 we have

$$\frac{-w_{12}}{\sqrt{w_{11}w_{22}}} = \frac{\frac{\rho\sigma_1\sigma_2}{\lambda}}{\sqrt{\frac{\sigma_1^2}{\lambda}\frac{\sigma_2^2}{\lambda}}} = \rho = \frac{u_{12}}{\sqrt{u_{11}u_{22}}}. \tag{4.5.24}$$

□

4.6 Testing the significance of conditional independence

Under Gaussianity, conditional dependence and partial correlation are equivalent.

Given a sample of observations of $\boldsymbol{X} \sim MVN(\mu, \boldsymbol{V})$, i.e. $x^1, \dots x^n$, we can form the usual estimation of $\boldsymbol{V}$ as

$$\hat{\boldsymbol{V}} = \frac{1}{n}\sum_i (x^i - \bar{x})(x^i - \bar{x})^T$$

and from $\hat{\boldsymbol{V}}$ construct $\hat{\boldsymbol{W}}$ and the sample partial correlations as in the inverse variance lemma. Hence in order to identify the edge set of a CIG we can test if the sample partial correlations, so obtained, are significantly different from zero. There is a useful relationship between the test of the partial correlation and the test for a regression coefficient. Let β_{ij} be the coefficient of x_j in the regression of x_i on x_j and X_c, then

$$\beta_{ij} = 0 \Longleftrightarrow \tau_{ij} = 0. \tag{4.6.1}$$

Let this regression be carried out using the sample $x^1, \ldots x^n$ of $\boldsymbol{X}$ and let t_{ij} be the t coefficient of the estimated parameter β_{ij}. Then the sample partial correlation τ_{ij} is related to the value of t_{ij} by (see Greene, 1993, p. 180)

$$\tau^2 = \frac{t^2}{t^2 + \nu} \tag{4.6.2}$$

where $\nu = n - k$ are the residual degrees of freedom of the regression. From this we can establish critical values for τ from the critical values of t. An alternative approach, by log-likelihood ratio, has been proposed by Whittaker (1990, p. 189). He suggests using the asymptotic properties of maximum likelihood estimates which leads to an asymptotically equivalent large sample distribution:

$$-n\log(1 - \tau^2) \sim \chi^2_1. \tag{4.6.3}$$

These tests strictly apply only to testing a single partial correlation. In practice we shall use them to "screen" all the partial correlations to identify a preliminary CIG structure. This is similar to the use of the partial autocorrelation function for identifying AR model order in time series. The graphs so identified will be used to formulate the model which will be then fitted and tested rigorously. When the variables we are dealing with are not Gaussian the null hypothesis of the test is not independence anymore but the lack of predictability.

4.7 An application to innovations in interest rates

In this section we apply the methods described to analyse the causal relations between residuals. This analysis has been pursued by Tunnicliffe Wilson in 1992. A similar methodology has been recently developed, on a different dataset, by Swanson and Granger (1997). The investigation of structural relations among residuals has important implications in the calculation of *impulse response functions* (IRF's) and *forecast error variance decompositions* (FEVD's).

Tunnicliffe Wilson investigated the relation between seven different maturities of the U.S. dollar interest rate: 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years (figure 4.17). Such a relation is known in the economic literature as *term structure*. The dataset consists of seven time series of 650 daily observations over the period 30^{th} November 1987-4^{th} December 1990.

He estimated a MARMA(1,1) for these time series and then used graphical modelling to analyse the structural relation among the residuals (figure 4.18). First, he computed the sample error correlation matrix, which is displayed in table 4.1 and then the scaled

Table 4.1: Correlation coefficients of model residuals.

	column						
row	1	2	3	4	5	6	7
1	1.00						
2	0.80	1.00					
3	0.50	0.53	1.00				
4	0.47	0.50	0.94	1.00			
5	0.43	0.45	0.88	0.92	1.00		
6	0.39	0.41	0.84	0.89	0.96	1.00	
7	0.40	0.41	0.81	0.86	0.94	0.97	1.00

inverse correlation matrix (table 4.2) which, according to the inverse variance lemma, gives the partial autocorrelations.

Table 4.2: Inverse correlation coefficients of model residuals with * indicating significance at the 2% level.

	column						
row	1	2	3	4	5	6	7
1	1.00						
2	*-0.73	1.00					
3	-0.04	*-0.10	1.00				
4	0.00	-0.03	*-0.69	1.00			
5	-0.06	0.04	*-0.16	*-0.25	1.00		
6	0.07	0.01	0.04	-0.05	*-0.54	1.00	
7	-0.04	-0.01	0.06	-0.02	*-0.10	*-0.67	1.00

The significant partial autocorrelations are indicated by a $*$ on the side. To test the significance of such partial correlations Tunnicliffe Wilson used the threshold $\pm 2/\sqrt{n}$ designed for individual test when inspecting sample partial autocorrelations of time series. However the results have been confirmed by the tests described above. The resulting CIG is shown in fig. 4.19

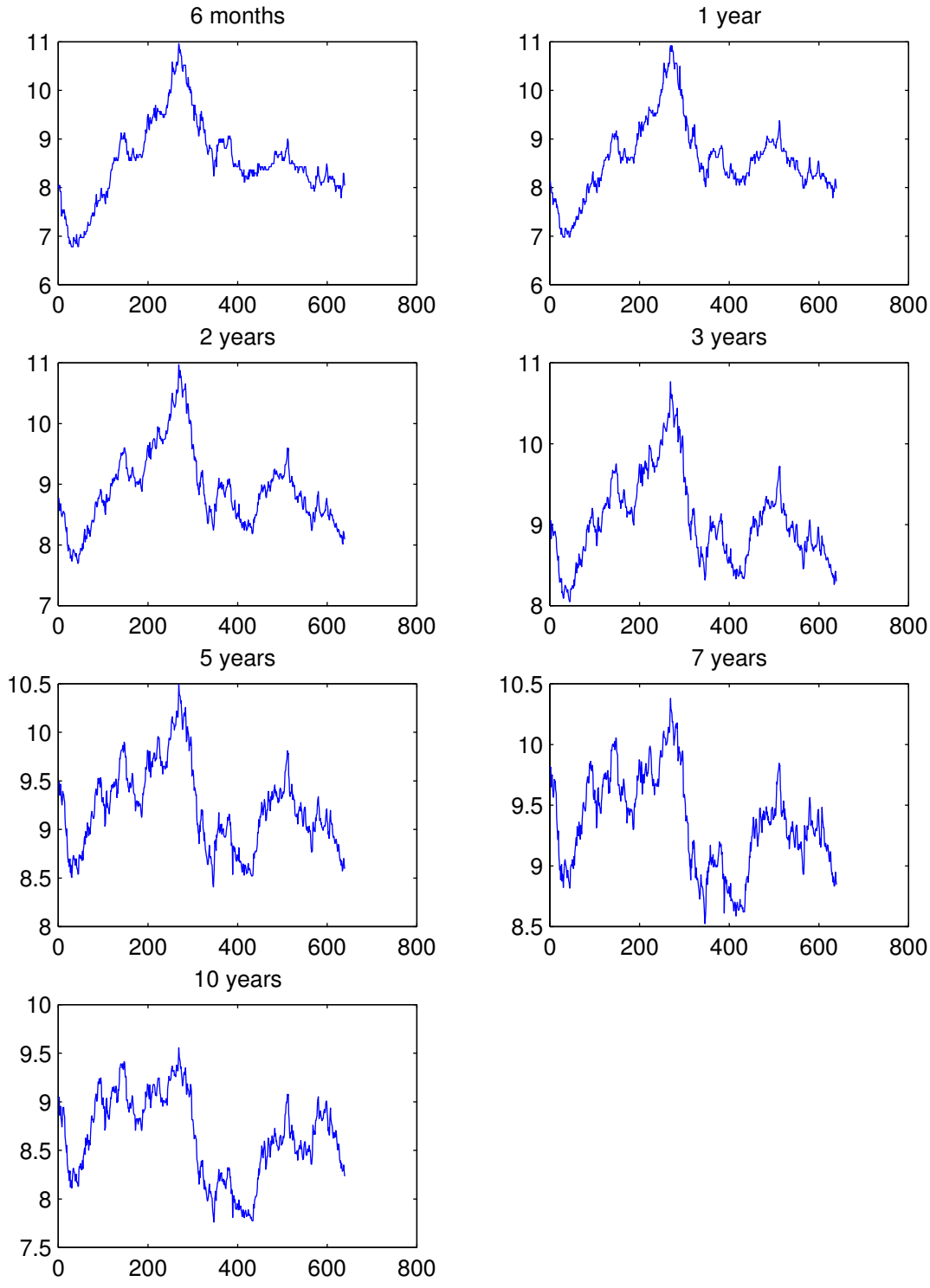


Figure 4.17: Different maturities of the U.S. dollar interest rate.

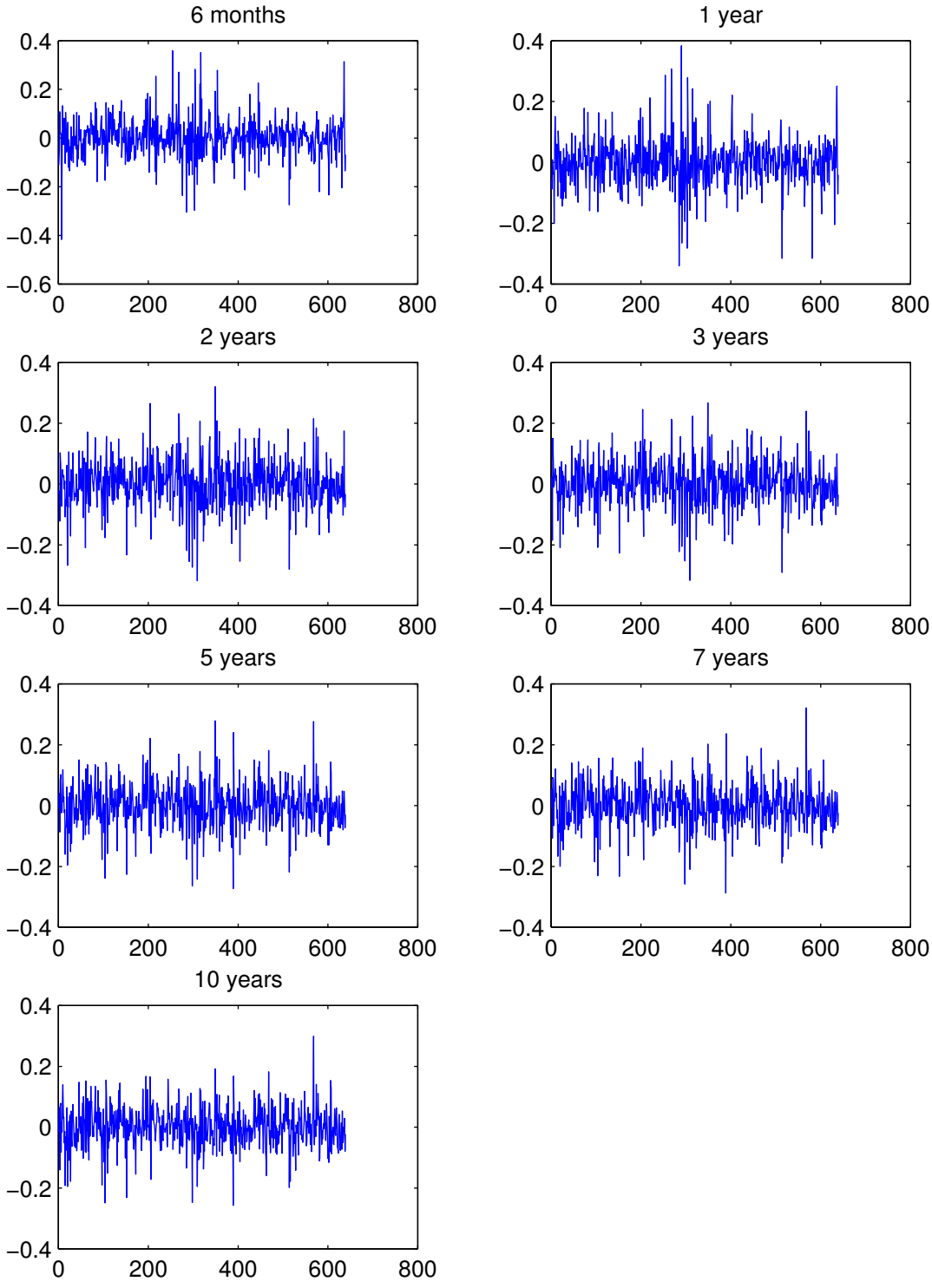


Figure 4.18: Residuals from a MARMA(1,1) model for the U.S. dollar interest rate maturities.

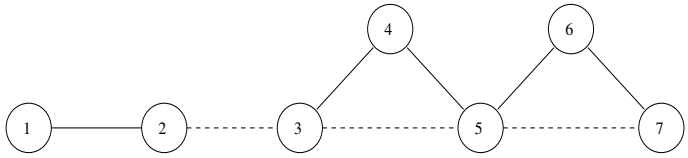


Figure 4.19: CIG for the term structure of the U.S. interest rate.

where the dotted lines represent a weaker significance of the test.

4.8 Gaussian DAG models

A DAG describes the marginal dependence of the variables. This is expressed in general as a factorisation of the joint density function (see equation 4.4.13).

In the Gaussian context each marginal dependence term is simply a regression equation for a variable, using as regressors those variables which are directed 'causal' variables.

The likelihood of a model or rather its deviance (-2 log-likelihood)is therefore the sum of the deviance contributions from each of the marginal terms. For the Gaussian model we take the deviance for each term to be ln S where S is the residual sum of squares from the regression. For example, if the DAG is the graph in figure 4.20 its density function

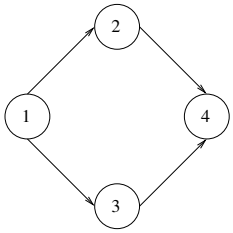


Figure 4.20: DAG.

is

$$f_{1,2,3,4}(\cdot) = f_{4|\{2,3\}}(\cdot) f_{3|1}(\cdot) f_{2|1}(\cdot) f_1(\cdot)$$

so the overall deviance is

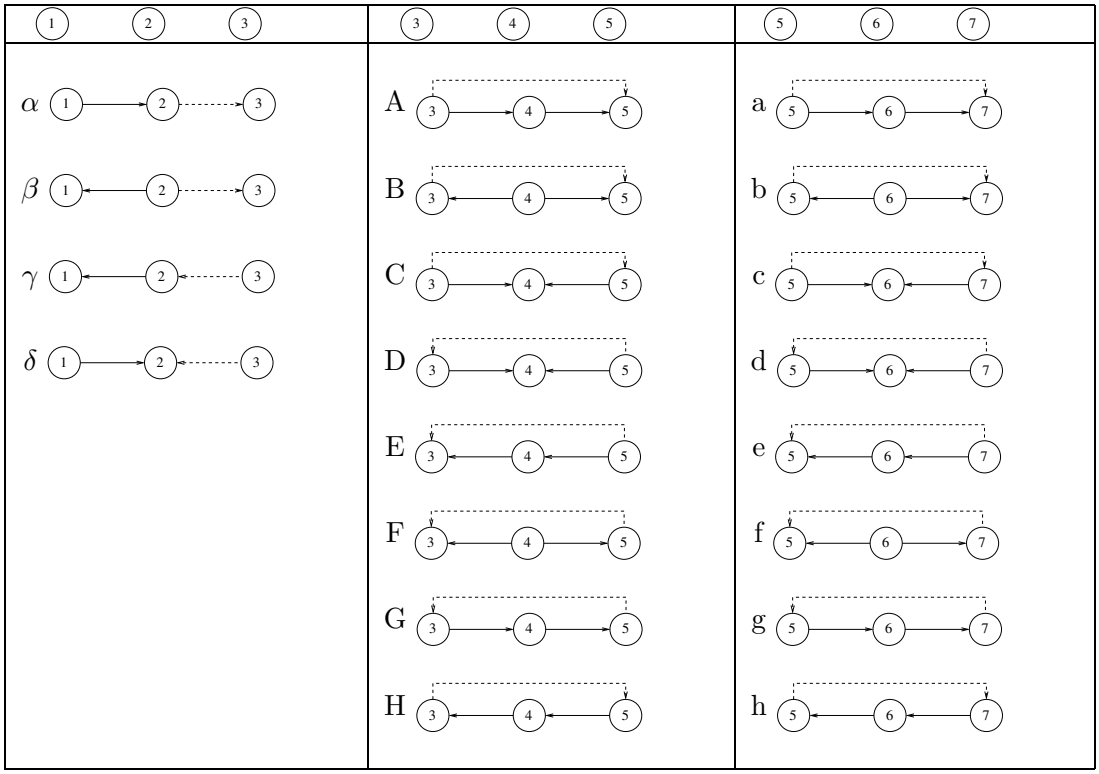
$$\text{Dev} = \ln S_4 + \ln S_3 + \ln S_2 + \ln S_1 \tag{4.8.1}$$

where S_4 is the sum of the squares from the regression of X_4 on X_3 and X_2, S_3 from the regression of X_3 on X_1, similarly for S_2 while S_1 is the raw sum of squares for X_1.

4.9 Further application to interest rates

We can continue the analysis of section 4.7 and try to model the causal structure between the variables. Our strategy was to attach directions to the seven edges and then to moralise the resulting DAG, to check whether it gave the observed CIG. Note that for 7 variables there are $\frac{7\times6}{2} = 21$ edges and 2^{21} possible directed graphs. For the identified CIG with just 8 edges this reduces to $2^8 = 256$ directed graphs but many or them are cyclic or inconsistent with the CIG in figure 4.19. To find out the possible graphs, we have decomposed the CIG in three subgraphs and then listed all the possibilities for each subgraph. They are presented in table 4.3.

Table 4.3: Possible directed subgraph.



We can obtain all the possible directed graphs by combining the possible subgraphs; in fact they are $4 \times 8 \times 8 = 256 = 2^8$. As an example if we combine α with C and a, we obtain the graph in figure 4.21.

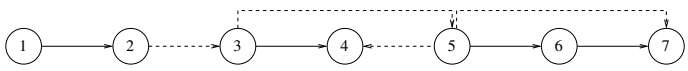


Figure 4.21: Graph αCa.

We can immediately exclude all the combinations containing F,G,f and g because they are cyclic. Also we can exclude all the combinations containing δ as it is inconsistent with the CIG in figure 4.19 because links 1 and 3, pa(2), are not married. Hence we have already reduced the number of possible graph to $3 \times 6 \times 6 = 108$ and we can still reduce their number as some of the combinations are inconsistent with the original CIG because of the lack of moral links. Eventually there are just 28 possible combinations: 1)αAa; 2)αAc; 3)αCa; 4)αCc; 5)βAa; 6)βAc; 7)βCa; 8)βCc; 9)γAa; 10)γAc; 11)γBa; 12)γBc; 13)γCa; 14)γCc; 15)γDa; 16)γDb; 17)γDc; 18)γDd; 19)γDe; 20)γDf; 21)γEa; 22)γEb; 23)γEc; 24)γEd; 25)γEe; 26)γEf; 27)γFa; 28)γFc.

All these models have the same likelihood; this is because:

1) the densities of the allowed subgraphs $\alpha,\ \beta,\ \gamma$ are the same and similarly so are those of the subgraphs A to F and those of a to f, e.g. for α

$$f_{1,2,3}(\cdot) = f_1(\cdot) f_{2|1}(\cdot) f_{3|\{2,1\}}(\cdot)$$

and because the Markov's properties of the original CIG

$$f_{1,2,3}(\cdot) = f_1(\cdot) f_{2|1}(\cdot) f_{3|2}(\cdot).$$

For β

$$\begin{aligned} f_{1,2,3}(\cdot) &= f_{\{1,3|2\}}(\cdot) f_2(\cdot) \\ &= f_{1|2}(\cdot) f_{3|2}(\cdot) f_2(\cdot); \end{aligned} \tag{4.9.1}$$

2) on combining these subgraphs no new links are formed by moralisation, so that no variable at an intersection (i.e. 3 or 5) is caused by variables in both subgraphs. This allows to factorise the complete density into those associated with the subgraphs, e.g.

$$f_{1,2,3,4,5,6,7}(\cdot) = f_{1,2,3}(\cdot) f_{4,5|1,2,3}(\cdot) f_{6,7|1,2,3,4,5}$$

which by the Markovian properties of the CIG becomes

$$f_{1,2,3,4,5,6,7}(\cdot) = f_{1,2,3}(\cdot) f_{4,5|3}(\cdot) f_{6,7|5}$$

which includes the density described by, for example the DAG αAa.

Because of the Gaussianity, the deviance for all the models can be computed according to 4.8.1, so for example the deviance of αAa model, whose joint density function is

$$f_{1,2,3,4,5,6,7}(\cdot) = f_1(\cdot) f_{2|1}(\cdot) f_{3|2}(\cdot) f_{4|3}(\cdot) f_{5|\{4,3\}}(\cdot) f_{6|5}(\cdot) f_{7|\{5,6\}}(\cdot),$$

is explained by the sum of squares of the residuals from the regressions of: 1 on no variables, hence we take the raw sum of squares of the observations; 2 on 1; 3 on 2, etc.... Note that there are just 20 distinct regressions whose sum of squares can be used to form the deviances of the 28 possible models.

As all the models consist of the same likelihood and have the same number of parameters, we cannot use likelihood based methods to select one and at this stage any prior information about the model should be used.

To further restrict the number of alternative possible models we can follow two different strategies: choosing the most parsimonious model or the explicative one. If we go for the first strategy we have to check which are the least significant links and take them out. In this way we obtain models with a smaller number of parameters without great loss in terms of likelihood. To check the significance we examine the t-values from the single regression performed to compute the deviance. In our case the weakest links appears to be the ones between: variables 3 and 5 in the regression of 3 on 4 and 5 and in the regression of 5 on 4 and 3; the link between variables 5 and 7 in the regression of 5 on 6 and 7 and in the regression of 7 on 5 and 6. Consistently with this strategy we then select the models where such links, possibly both, are present and take them out. Hence from this point of view we should select: αAa'; βAa'; γAa'; γBa'; γEa'; γEb'; γEe' and γEf' where the dash implies that we have taken out the weak links from the original graphs. We can observe that γEb' and γEf' are identical and this eventually leads to seven possible graphs which are presented in figure 4.22. These models can represent adequately the dynamics within the residuals of our original time series with a comparatively small number of parameters. This is a very useful feature if we are interested in the forecasting capability. Nevertheless they may not be very enlightening if we wish to explain the dynamics of the series through time.

According to the economic theory, in the financial markets there are some interest rates whose movements affects with a cascade mechanism all the others, such interest rates are indicated as *pivot* interest rates. In this respect our selection is almost vague as any variable could be the pivot interest rate: 1 in model αAa'; 2 in model βAa'; 3 in model γAa'; 4 in model γBa'; 5 in model γEa'; 6 in models γEb and γEf; 7 in model γEf.

To decide to which extent we can take out links, we can use likelihood based methods such as AIC, BIC, etc... which take the number of parameters into account and which will be described later.

Figure (4.22.1): αAa'

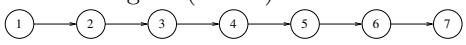


Figure (4.22.2): βAa'

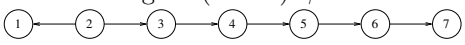


Figure (4.22.3): γAa'

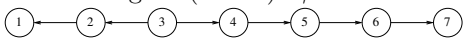


Figure (4.22.4): γBa'

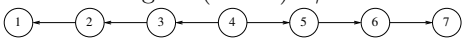


Figure (4.22.5): γEa'

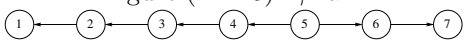


Figure (4.22.6): γEb' and γEf'

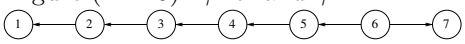


Figure (4.22.7): γEe'

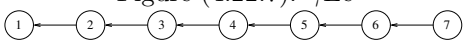


Figure 4.22: Possible DAG's selected using the strategy of the most parsimonious model (MPM).

The alternative strategy is to select the most explicative models (MEM), that is the models where the highest possible number of links are present, i.e. where the least number of weak links above identified are present. Following this strategy we would select models: αCc; βCc; γCc; γDc and γDd. This strategy may give a clearer picture of the dynamics. In our case, despite of the fact that the selection has left a considerable number of alternative models, shown in figure 4.23, it gives a more restricted representation of the structural causation among the interest rates. These five models support the existence

Figure (4.23.1): αCc

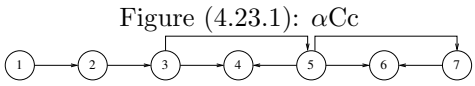


Figure (4.23.2): βCc

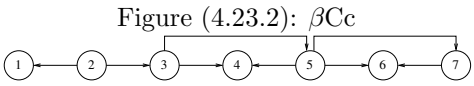


Figure (4.23.3): γCc

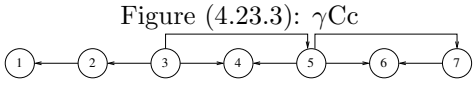


Figure (4.23.4): γDc

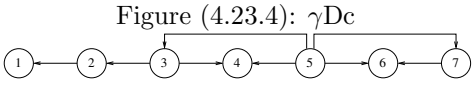


Figure (4.23.5): γDd

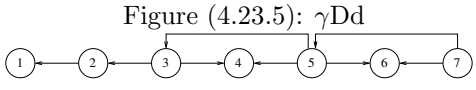


Figure 4.23: Possible DAG's selected using the strategy of the most explicative model (MEM).

of only five alternative pivot interest rates: variable 1 (model αCc); 2 (βCc); 3 (γCc); 5 (γDc) and variable 7 (models γDd).

In order to give a better representation of the model dynamics we can attach values to links which quantify the influence. Such values can be obtained by the regressions performed to compute the deviance. For example model γBa can be represented as in figure 4.24.

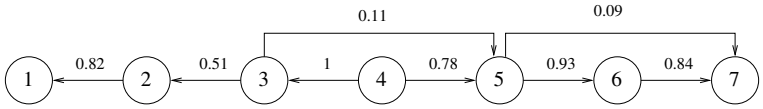


Figure 4.24: Graph γBa with link values.

A way to reduce the number of possible models is to use higher thresholds for the significance test of the partial correlations but, as usual, this leads to the trade-off between type I and type II errors. In our application, using a stricter test would mean taking out the dotted edges in the original CIG with a subsequent separation of the first two variables. This is consistent with the economic theory as it represents a division between the monetary market (maturity within a year) and the financial market (term to maturity longer than a year). This has its rationale in the importance of the transactive component in the money demand function for the monetary market. Such a component is almost absent in the financial market and implies independent dynamics for the two markets.

In chapter 6 we shall see how including the time regression structure clarifies the choice of pivot variable as the 2 year interest rate.

4.10 Diagnostic checking

The basic assumption underlying the Gaussian DAG models is that the regression error variables are mutually uncorrelated. This may be checked by inspecting the sample correlation matrix of the residuals. If this shows evidence of any remaining significant correlation, the assumed DAG must be amended. This may occur because the original identification of the DAG using the partial correlation matrix is only "tentative". For the model *gammaCc* of the previous section, the error correlation matrix,$\boldsymbol{R}$, is represented in table 4.4. The element $r_{1,3}$ of $\boldsymbol{R}$ is well in excess of the appropriate standard error limit of approximately $\frac{2}{\sqrt{n}} = 0.08$. We therefore explore possible model extensions, first by introducing just one further link, 1-3, as in figure 4.25. The correspondent residual correlation matrix is represented in table 4.5 and the correspondent likelihood is -2.38.

Table 4.4: Residual correlation matrix.

1	1.00						
2	-0.08	1.00					
3	0.13	0.00	1.00				
4	-0.01	0.04	0.00	1.00			
5	0.03	-0.04	0.00	0.00	1.00		
6	-0.06	-0.06	-0.05	0.08	0.09	1.00	
7	0.03	0.06	-0.02	0.02	0.04	0.00	1.00
	1	2	3	4	5	6	7

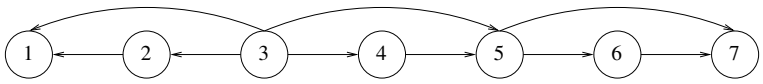


Figure 4.25: Graph γCa with an added link.

Table 4.5: Residual correlation matrix of the modified model.

1	1.00						
2	-0.00	1.00					
3	-0.00	-0.00	1.00				
4	-0.01	0.04	0.00	1.00			
5	0.03	-0.04	0.00	0.00	1.00		
6	-0.06	-0.06	-0.05	0.08	0.09	1.00	
7	0.04	0.06	-0.02	0.02	0.04	0.00	1.00
	1	2	3	4	5	6	7

This is much improved (see table 4.6) and so we accept the model. A further check is to compare the likelihood with that of the saturate model which is -2.41. Our final model is clearly preferred using AIC, SIC or HIC criteria, which will be described in chapter 5.

Table 4.6: Comparisons of the different models.

model	par.	lik	AIC	SIC	HIC
saturated	21	-2.41	-1500.3	-1406.6	-1271
γCa	8	-2.35	-1488.7	-1453	-1401.4
fig. 4.25	9	-2.38	-1500.8	-1460.7	-1402.6

Chapter 5

Graphical Modelling Approach to Univariate AR Models

This chapter links the concept of graphical modelling to time series in the univariate context. It shows how GM can aid the understanding and the identification of univariate AR models. Its potential is limited in this context, being realized to much greater extent with the multivariate models of the following chapter. However, some of the ideas are usefully introduced at this stage. We can therefore consider a finite selection of values $X_t, X_{t-1}, \ldots, X_{t-k}$ of a stationary time series to be variables to which we apply graphical modelling methods. Under Gaussianity conditions we can estimate the covariance matrix $\boldsymbol{V}$ of $X_t, X_{t-1}, \ldots, X_{t-k}$ and then derive the matrix of the estimated partial correlations. We need to understand the structure which we expect to find when we carry on with this procedure, i.e. the structure of the CIG described by the partial correlations. In particular we are interested in the structure we should find when $\{X_t\}$ follows an AR(p) model with $p < k$. We work out the theoretical structure and also consider the statistical issues related to the significance test for the edges in the CIG. The distribution of the test statistics cannot be assumed to be the same, in a time series, as that used in the previous chapter for data from IID MVN(0,$\boldsymbol{V}$).

5.1 The DAG and CIG structure of the AR(p) model

Consider the observations $x_1, \ldots, x_n$ where $n > p$. Assuming, at this stage, no special structure for the first p values, the AR(p) model prescribes the dependence, i.e. the DAG structure, relating successive values.

Example 5.1.1 *(AR(1)). The DAG for a AR(1) model is represented by the graph in figure 5.1 giving the CIG to be exactly the same with the directions removed.* □

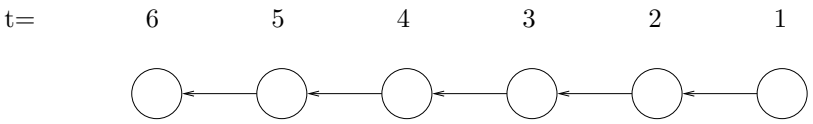


Figure 5.1: DAG of a AR(1) model

Example 5.1.2 *(AR(1)). The DAG for a AR(3) model is represented by the graph in figure 5.2.*

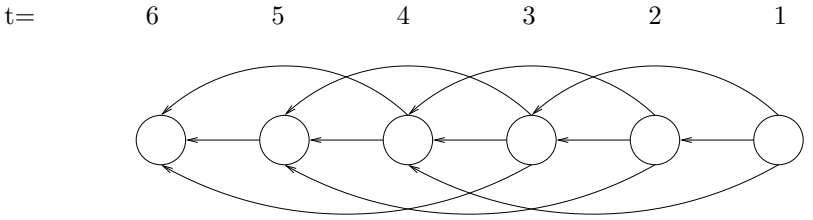


Figure 5.2: DAG of a AR(3) model

Note that a general complete dependence is assumed for x_1, x_2, x_3. The resulting CIG is shown in figure 5.3 □

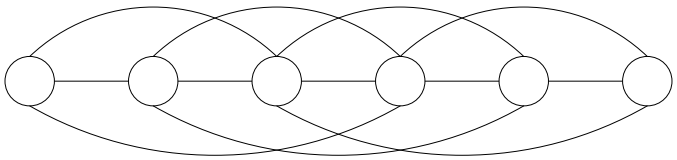


Figure 5.3: Moral graph for a AR(3) model

Example 5.1.3 *(Subset AR(3) model). Consider a subset AR(3) model with $\phi_2 = 0$, its DAG is shown in figure 5.4 with an arbitrary initial distribution for x_1, x_2, x_3.*

Moralisation gives the CIG of figure 5.5.

So in the CIG there is only one link which, by its absence, reflects the subset structure. If, however, the initial distribution of x_1, x_2, x_3 is not arbitrary, because of the stationary distribution of the process, all the graphs must reflect the time reversibility of the autocovariance structure of $\{X_t\}$, in this case we will also find another link absent, as shown in figure 5.6.

□

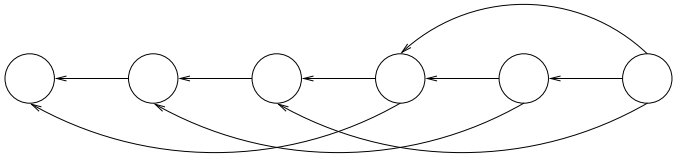


Figure 5.4: DAG of a subset AR(3) model

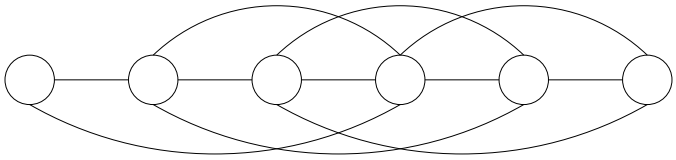


Figure 5.5: Moral graph for a subset AR(3) model with $\phi_2 = 0$

5.2 The CIG structure of the stationary AR(p) process

Consider the AR(p) model

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + e_t$$

and define the $n \times n$ and $p \times n$ matrices

$$\boldsymbol{M}_n = \begin{bmatrix} 1 & -\phi_1 & \ldots & -\phi_p & 0 & \ldots & 0 \\ 0 & 1 & \ldots & -\phi_{p-1} & -\phi_p & \ldots & 0 \\ 0 & 0 & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & -\phi_p \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \ldots & 0 & \ldots & 1 \end{bmatrix} \tag{5.2.1}$$

and

$$\boldsymbol{N}_n = \begin{bmatrix} 0 & \ldots & -\phi_p & \ldots & \ldots & -\phi_2 & -\phi_1 \\ 0 & \ldots & 0 & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & \ldots & \ldots & -\phi_p & -\phi_{p-1} \\ 0 & \ldots & 0 & \ldots & \ldots & 0 & -\phi_p \end{bmatrix} \tag{5.2.2}$$

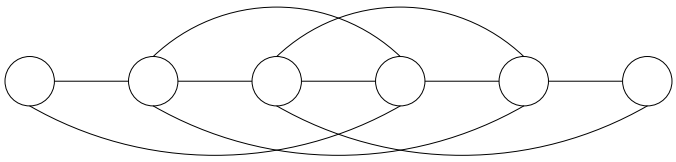


Figure 5.6: Moral graph for a subset AR(3) model with $\phi_2 = 0$ including symmetry implied by time reversibility of the stationary process.

and let $\boldsymbol{V}_n = \mathrm{Var}\,(x_1, \ldots, x_n)$, then (see Ljung and Box, 1979), assuming that $(x_1, \ldots, x_p)$ come from the stationary distribution

$$\sigma_e^2 V^{-1} = M^T M - N^T N. \tag{5.2.3}$$

Consider, now, the partitioned matrix

$$R = \begin{bmatrix} 1 & -\phi_1 & \ldots & -\phi_p & 0 & \ldots & 0 \\ 0 & 1 & \ldots & -\phi_{p-1} & -\phi_p & \ldots & 0 \\ 0 & 0 & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & 1 & \vdots & \vdots & -\phi_p \\ \multicolumn{7}{c}{\dotfill} \\ \multicolumn{3}{c}{\mathbf{0}} & \vdots & \multicolumn{3}{c}{\boldsymbol{I}_p} \end{bmatrix} = \begin{bmatrix} \multicolumn{3}{c}{\boldsymbol{M}_{n-p,n}} \\ \multicolumn{3}{c}{\dotfill} \\ \mathbf{0} & \vdots & \boldsymbol{I}_p \end{bmatrix} \tag{5.2.4}$$

and the case when $(x_1, \ldots, x_p)$ are taken from an *arbitrary* MVN distribution.

Then

$$\boldsymbol{R} \begin{bmatrix} x_n \\ \vdots \\ x_{p+1} \\ x_p \\ \vdots \\ x_1 \end{bmatrix} = \begin{bmatrix} e_n \\ \vdots \\ e_{p+1} \\ x_p \\ \vdots \\ x_1 \end{bmatrix} = \boldsymbol{Y} \tag{5.2.5}$$

where $\boldsymbol{Y}$ has covariance matrix

$$\begin{bmatrix} \sigma_e^2 \boldsymbol{I}_{n-p} & \vdots & \mathbf{0} \\ \multicolumn{3}{c}{\dotfill} \\ 0 & \vdots & \boldsymbol{V}_p \end{bmatrix} = \boldsymbol{U}. \tag{5.2.6}$$

Then, from 5.2.5

$$\boldsymbol{R}\boldsymbol{V}_n\boldsymbol{R}^T = \boldsymbol{U} \tag{5.2.7}$$

and

$$\sigma_e^2 \boldsymbol{V}_n^{-1} = \boldsymbol{R}^T \boldsymbol{U}^{-1} \boldsymbol{R} = \boldsymbol{M}_{n-p,n}^T \boldsymbol{M}_{n-p,n} + \begin{bmatrix} 0 & \vdots & 0 \\ \multicolumn{3}{c}{\dotfill} \\ 0 & \vdots & * \end{bmatrix} \tag{5.2.8}$$

where the asterisk indicates part of the matrix in which we are not interested. As a consequence of 5.2.8 all the links between $x_n, \ldots, x_1$ except those between $x_p \ldots x_1$ can be determined by examining $\boldsymbol{M}_{n-p}^T \boldsymbol{M}_{n-p}$. This corresponds exactly to the moralisation of the DAG describing the model.

In the case when the process is stationary, $\boldsymbol{V}_n$ has time reversal symmetry. This is reflected in the formula 5.2.3 which extends 5.2.8 to

$$\sigma^2 \boldsymbol{V}_n^{-1} = \boldsymbol{M}_n^T \boldsymbol{M}_n - \boldsymbol{N}_n^T \boldsymbol{N}_n. \tag{5.2.9}$$

This is identical to 5.2.8 except for the last $p \times p$ submatrix.

Example 5.2.1 *(AR(2)).*

$$\sigma^2 \boldsymbol{V}_n^{-1} = \begin{bmatrix} 1 & -\phi_1 & -\phi_2 & & & \\ -\phi_1 & 1+\phi_1^2 & \phi_1\phi_2 - \phi_1 & & * & \\ & & 1+\phi_1^2+\phi_2^2 & & & \\ & & & \ddots & & \\ & * & & & 1+\phi_1^2 & -\phi_1 \\ & & & & -\phi_1 & 1 \end{bmatrix} \tag{5.2.10}$$

for $n > 2$. □

Note first that apart from the first and last $(p-1)\times(p-1)$ diagonal submatrices, all the entries are constant down any diagonal. The first $(p-1)\times(p-1)$ submatrix is different because of the *end effect* upon moralisation. The last $(p-1)\times(p-1)$ reflects this by time reversal symmetry of stationary autocorrelation.

Note also that the first row of $\boldsymbol{V}_n^{-1}$ or equivalent of $\boldsymbol{W}_n$, the matrix of partial correlations, corresponds to the coefficients in the AR(p) model up to lag p and zero thereafter. The last coefficient in the first row corresponds to Corr $(X_n, X_1 \mid X_{n-1}, \ldots, X_2)$ which is the partial correlation at lag (n-1) of the series.

The question we ask is whether this matrix of partial correlations, or just its first row is of any value in addition to the standard partial autocorrelation function of the series.

The standard partial a.c.f. is generally used to identify the order of the model. This is equivalent to fitting increasing orders of AR(p) model and testing for the coefficient ϕ_p in each model.

The use of the CIG approach by directly estimating the partial correlations for some large but fixed size k, is equivalent to fitting an AR(k) model, for $k > p$, and examining the individual coefficients $\phi_1, \ldots, \phi_k$ in this model.

One way to compare the approaches is to compare the power of testing $\phi_p = 0$ when the order is p (i.e. the true value of the ϕ_p is not zero but all the coefficient $\phi_j,\ j > p$ are zero) by:

1) fitting the AR(p) model and testing $\phi_p = 0$;

2) fitting the AR(k) for some $k > p$ and testing $\phi_p = 0$.

This comes down to compare the standard deviation of ϕ_p estimated by these two methods. To answer this question we use the fact that in large samples the variance matrix

of the coefficients of a AR(k) model is (see Box and Jenkins, 1976)

$$\frac{1}{n}\sigma_e^2 \boldsymbol{V}_k^{-1} = \frac{1}{n}(\boldsymbol{M}_k^T \boldsymbol{M}_k - \boldsymbol{N}_k^T \boldsymbol{N}_k) \tag{5.2.11}$$

For example if p=1 and k=2, fitting an AR(1) the standard error is obtained by

$$\frac{1}{n}\sigma_e^2 \boldsymbol{V}_1^{-1} = \frac{1}{n}\left[(1)^T(1) - (-\phi_1)^T(-\phi_1)\right] = \frac{1}{n}\left[(1-\phi_1^2)\right] \tag{5.2.12}$$

i.e.

$$\text{Var}\,(\hat{\phi}_1) = \frac{1}{n}(1-\phi_1^2); \tag{5.2.13}$$

but fitting an AR(2),

$$\sigma_e^2 \boldsymbol{V}_2^{-1} = \begin{bmatrix} 1 & -\phi_1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & -\phi_1 \\ 0 & 1 \end{bmatrix}^T - \begin{bmatrix} -\phi_2 & -\phi_1 \\ 0 & -\phi_2 \end{bmatrix}\begin{bmatrix} -\phi_2 & -\phi_1 \\ 0 & -\phi_2 \end{bmatrix}^T \tag{5.2.14}$$

that is

$$\sigma_e^2 \boldsymbol{V}_2^{-1} = \begin{bmatrix} 1+\phi_1^2 & -\phi_1 \\ -\phi_1 & 1 \end{bmatrix} - \begin{bmatrix} \phi_1^2+\phi_2^2 & \phi_1\phi_2 \\ \phi_1\phi_2 & \phi_2^2 \end{bmatrix} \tag{5.2.15}$$

Then (see Box and Jenkins, 1976, Appendix 7.5.24)

$$\sigma_e^2 \boldsymbol{V}_2^{-1} = \begin{bmatrix} 1-\phi_2^2 & -\phi_1-\phi_1\phi_2 \\ -\phi-\phi_1\phi_2 & 1-\phi_2^2 \end{bmatrix}. \tag{5.2.16}$$

If $\phi_2 = 0$

$$\sigma_e^2 \boldsymbol{V}_2^{-1} = \begin{bmatrix} 1 & -\phi_1 \\ -\phi_1 & 1 \end{bmatrix}. \tag{5.2.17}$$

and so

$$\text{Var}\,(\hat{\phi}_1) = \frac{1}{n} \tag{5.2.18}$$

Similarly, if p=1 and k takes any value $k > 1$, then $\text{Var}\,(\hat{\phi}_1) = \frac{1}{n}$. This is also the 'default' variance of the sample partial autocorrelation function used in selecting the order by plotting the pacf within bounds of $\pm\frac{2}{\sqrt{n}}$. However, the standard errors of $\hat{\phi}_j$, $j = 2, \ldots, k-1$ are given, using 5.2.3, by

$$\text{SE}(\hat{\phi}_j) = \frac{1}{\sqrt{n}}(1+\phi_1^2)^{\frac{1}{2}}. \tag{5.2.19}$$

In general, on fitting an AR(p) model, using 5.2.3, we have

$$\text{SE}(\hat{\phi}_p) = \frac{1}{\sqrt{n}}(1-\phi_p^2)^{\frac{1}{2}} \tag{5.2.20}$$

whereas on fitting an AR(k) of order greater than p,

$$\text{SE}(\hat{\phi}_j) = \frac{1}{\sqrt{n}}(1+\phi_1^2+\ldots+\phi_p^2)^{\frac{1}{2}}. \tag{5.2.21}$$

Assuming that $\phi_p \neq 0$ the power of the test to reject $\phi_p = 0$ will be less for the second option.

The conclusion is that examining the first row of the sample partial correlation matrix will not provide as great a power for selecting the order of the AR(p) model, as would be obtained from the usual method of examining the sample partial autocorrelation function. On the other hand, it does allow simultaneous examination of all the estimates $\hat{\phi}_j$, whereas using the partial autocorrelation function it would be necessary first to select the order p and then to estimate and examine the parameters of the AR(p) model.

One of the advantages of estimating ϕ_j for a high order k model is that it might be possible to see a sparse structure in the coefficients, i.e. of $\phi_j = 0$ for j less than the supposed order p. Examination of the partial autocorrelation function does not generally reveal this sparse structure.

Example 5.2.2 . *For an AR(3) model*

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)x_t = 0 \tag{5.2.22}$$

the sample partial autocorrelation at lag 2 is

$$\phi_{2,2} = -\frac{\phi_2 - \phi_1\phi_3}{1 - \phi_3^2} \tag{5.2.23}$$

which is not zero even if ϕ_2 is zero. The sparse structure $\phi_2 = 0$ is not revealed by the value of $\phi_{2,2}$. □

Example 5.2.3 . *Consider the AR(5) model*

$$(1 - \phi_1 B)(1 - \phi_4 B^4)y_t = e_t \tag{5.2.24}$$

which factorizes the equation

$$y_t = \phi_1 y_{t-1} + \phi_4 y_{t-4} + \phi_5 y_{t-5} + e_t \tag{5.2.25}$$

where $\phi_5 = -\phi_1\phi_5$.

This is a realistic model for a seasonal behaviour of a quarterly time series similar, in the AR structure, to the multiplicative models of Box and Jenkins (1976, p. 303). In the generalised form

$$y_t = \sum_{j=1}^{5} \phi_j y_{t-j} + e_t$$

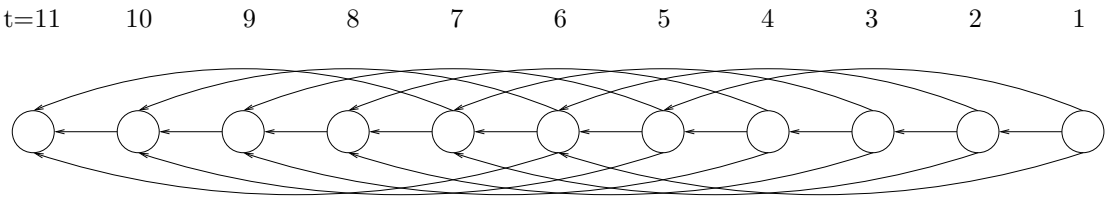


Figure 5.7: DAG for the AR(5) model with $\phi_2 = \phi_3 = 0$.

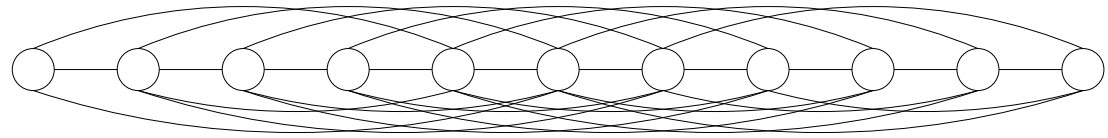


Figure 5.8: Moral graph for the AR(5) model with $\phi_2 = \phi_3 = 0$.

it is sparse as $\phi_2 = \phi_3 = 0$. *The corresponding DAG is presented in figure 5.7. Applying moralisation we obtain the CIG in figure 5.8 which is described by the partial autocorrealtion matrix in table 5.1 where* • *indicates a non-zero value. Note that, besides the*

	10	9	8	7	6	5	4	3	2	1
11	•	·	·	•	•	·	·	·	·	·
10		•	·	•	•	•	·	·	·	·
9			•	·	•	•	•	·	·	·
8				•	·	•	•	•	·	·
7					•	·	•	•	•	·
6						•	·	•	•	•
5							•	·	•	•
4								•	·	•
3									•	·
2										•

Table 5.1: Matrix of the significant theoretical partial autocorrelations.

zeros in the first row, which reflect the order 5 of the AR process and the zero coefficients at lags 2 and 3, there are also zeros reflecting absence of linkage in the lower rows. This may be useful as additional information in identifying the model. □

In the next subsection we will simulate an AR process with features similar to the example above and we apply graphical modelling techniques on the simulated time series in order to obtain the original structure.

5.2.1 A simulated example

We simulated 950 observations from the AR(5) model

$$(1 - 0.5\mathrm{B}^4)(1 - 0.8\mathrm{B})x_t = e_t \tag{5.2.26}$$

which corresponds to the equation

$$x_t = 0.8x_{t-1} + 0.5x_{t-4} - 0.4x_{t-5} + e_t. \tag{5.2.27}$$

Using the inverse variance lemma we computed the sample partial correlation matrix of the current and lagged values, presented in the table below

11	-0.65	0.02	-0.01	-0.33	0.32	-0.05	0.02	-0.06	0.01	0.04
10		-0.51	0.02	0.21	-0.44	0.28	-0.05	0.06	-0.05	0.01
9			-0.51	0.01	0.20	-0.47	0.30	-0.05	0.06	-0.06
8				-0.48	0.01	0.24	-0.52	0.30	-0.05	0.02
7					-0.53	0.02	0.24	-0.47	0.28	-0.05
6						-0.53	0.00	0.21	-0.44	0.32
5							-0.47	0.01	0.21	-0.33
4								-0.51	0.02	-0.01
3									-0.51	0.02
2										-0.65
	10	9	8	7	6	5	4	3	2	1

Applying the test as in 4.6.2 we obtain a threshold, for 0.99 significance, of about 0.08; according to this threshold we obtain the matrix in table 5.2 This matrix, representing

11	●	·	·	●	●	·	·	·	·	·
10		●	·	●	●	●	·	·	·	·
9			●	·	●	●	●	·	·	·
8				●	·	●	●	●	·	·
7					●	·	●	●	●	·
6						●	·	●	●	●
5							●	·	●	●
4								●	·	·
3									●	·
2										●
	10	9	8	7	6	5	4	3	2	1

Table 5.2: Matrix of the significant sample partial autocorrelations.

the significance of the sample partial correlations is different from the significance matrix of theoretical partial correlations of the previous example as the link between x_{t-7} and x_{t-10} (4 and 1) is not significant. This is because the moralisation in example 5.2.3 did

not consider the time reversal symmetry. The stationarity of the example does reveal this symmetry.

We also computed, for this model, the matrix of the partial autocorrelations of which the sample values are shown in the example in the subsection 5.2.1. We used formula 5.2.9 which is applicable to the stationary process. The pattern of non-zero elements is exactly as in table 5.2.

5.3 Sample properties of the matrix of partial autocorrelations

The testing procedure in the simulated example is 'borrowed' from the case of IID samples from the MVN distribution. We remarked at the beginning of this chapter that we cannot necessarily assume that this procedure can be applied in the time series context. In this section we demonstrate that this test is in fact justified in large samples for the first row of the partial autocorrelation matrix of the time series.

In section 4.6 we cited the relationship between the sample partial correlation τ and the regression t-statistics:

$$\tau^2 = \frac{t^2}{t^2 + \nu}.$$

This relationship remains valid in the time series context because it is an algebraic consequence of the relationship between the least squares equations and the computation of the partial correlation, both of which derive from the sample covariance matrix $\boldsymbol{V}$.

In the context of IID MVN data, the t-statistic under the usual Gaussianity assumptions, has the exact t distribution. In the time series, i.e. in the regression of x_t on $x_{t-1},\ x_{t-2}, \ldots, x_{t-k}$ using the sample values of $x_1, \ldots, x_k$ of the series, the t-statistic for the coefficient ϕ_j of x_{t-j} no longer follows an exact t distribution, because the rows of data $x_t, x_{t-1}, \ldots, x_{t-k}$, for $t = 1 + k, \ldots, n$, are no longer independent for different values of t: they come from the same series and are correlated.

For large n the t distribution on $(n - k)$ degrees of freedom approaches the standard normal distribution. In the time series context, it may also be shown for large n that the t statistic for the autoregressive coefficient ϕ_j approaches the standard normal distribution, under quite broad assumptions.

Provided, therefore, that the t distribution is replaced by the standard normal, we can,

in large samples, continue to use critical values based upon the above relationships for the first row of the sample partial correlation matrix of a time series when testing their significance. The proof of this sampling property for the autoregressive coefficients is given by theorems in Anderson (1971, p. 211–223). To summarise these, it is shown that in large samples

$$\hat{\phi}_j \sim MVN \quad \text{with} \quad \mathrm{E}(\hat{\phi}_j) \simeq \phi_j \qquad j = 1, \ldots, n \tag{5.3.1}$$

and

$$n\mathrm{Var}\,(\hat{\phi}_j) \simeq \boldsymbol{V}_k^{-1} \qquad j = 1, \ldots, n. \tag{5.3.2}$$

When the autoregressions are estimated by least squares, the regression vector and matrix are

$$\boldsymbol{Y} = \begin{bmatrix} x_n \\ \vdots \\ x_{k+1} \end{bmatrix} \quad \text{and} \quad \boldsymbol{X} = \begin{bmatrix} x_{n-1} & \ldots & x_{n-k} \\ \vdots & & \vdots \\ x_k & \ldots & x_1 \end{bmatrix}. \tag{5.3.3}$$

The least squares regression then gives the variance matrix of estimates as

$$\sigma_e^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \simeq \frac{1}{n}\boldsymbol{V}_k^{-1} \tag{5.3.4}$$

(assuming all vectors are mean corrected) because

$$\frac{1}{n}(\boldsymbol{X}^T\boldsymbol{X})_{ij} = \frac{1}{n}\sum_{t=1+k}^{n} x_{t-i}x_{t-j} \simeq \boldsymbol{C}_{i-j} \to \gamma_{i-j} = (\boldsymbol{V}_k)_{ij}.$$

Using the covariance matrix from the autoregression therefore gives the correct asymptotic variance for $\hat{\phi}$. Consequently

$$\frac{\hat{\phi}_j - \phi_j}{\mathrm{SE}(\phi_j)} \sim \mathrm{N}(0,1).$$

To quote Anderson (1971, p. 211):

"... *thus for large samples we can treat them* (the autoregressive coefficient estimates) *as normally distributed and the procedures of ordinary regression analysis can be used.*"

Note that this result assumes that the series can be represented by an AR model of order $p \leq k$; it is not true for a general time series.

We can only apply this result to the first row of the partial correlation matrix because we do not know of a similar result when x_{t-j} is regressed on both future and past values, i.e. $x_t, \ldots, x_{t-j+1}, x_{t-j-1}, \ldots, x_{t-k}$. Without such a result we cannot

establish the same sampling property for the partial correlations between x_{t-j} and $x_t, \ldots, x_{t-j+1}, x_{t-j-1}, \ldots, x_{t-k}$, i.e. the elements of the j^{th} row of the sample partial correlations. This result cannot be expected to be true because, in general, in the regression of a time series on both future and past values, the regression errors do not form an uncorrelated time series.

5.4 Comparing adequacy of different models

If we compare different models, we expect that a model involving a greater number of parameters has a closer fit to the observed data. However, for reasons of parsimony, we prefer models with fewer parameters. In particular it is important not to select models with an unnecessarily large number of parameters. Adequate fit is needed. There are some algorithms which compare different models taking into account the number of parameters and penalising models with an unnecessarily large number of parameters. Probably the most popular is the *Akaike Information Criterion* (AIC), defined by

$$AIC = -2\log[\text{maximised likelihood}] + 2k \tag{5.4.1}$$

where k is the number of parameters.

Extending the reasoning in section 3.6.3 for a general ARMA model of order (p,q), we can derive its maximum likelihood as (see Priestley, 1981, p. 373)

$$\max \hat{L} = -\frac{n}{2}\log(\hat{\sigma}_\epsilon^2) - \frac{n}{2} \tag{5.4.2}$$

where $\hat{\sigma}_\epsilon^2$ is the maximum likelihood estimate of the residual variance and n is the number of observations. Ignoring the second term of the RHS of 5.4.2, which is constant, 5.4.1 becomes

$$AIC = n\log(\hat{\sigma}_\epsilon^2) + 2k. \tag{5.4.3}$$

We should choose the model for which AIC is minimised.

Shibata (1976) showed that the AIC tends to overestimate the number of necessary parameters. Corrections have been proposed to give new criteria and among them, one by Schwartz (1978),

$$\text{SCH} = n\log(\hat{\sigma}_\epsilon^2) + k\log(n) \tag{5.4.4}$$

and one by Hannan and Quinn (1979)

$$\text{HAN} = n\log(\hat{\sigma}_\epsilon^2) + 2k\log(n). \tag{5.4.5}$$

It is apparent that Hannan's criterion penalises more, for the number of parameters, than Schwartz's criterion. In our analyses we shall refer to 5.4.4 when comparing different models.

Chapter 6

Graphical Modelling Approach to Multivariate AR Models

In this chapter we extend the material of the previous chapter to multivariate time series. Again we take a finite selection of values $x_t,\ x_{t-1}, \ldots, x_{t-k}$ of a stationary time series to be the variables to which we apply graphical modelling. However, $\boldsymbol{X}_t$ is a vector of variables, so that we are now modelling both the multivariate structure of the components of $\boldsymbol{X}_t$ and the time series structure. Again, assuming Gaussianity, we can estimate the covariance matrix $\boldsymbol{V}$ of $\boldsymbol{X}_t, \ldots, \boldsymbol{X}_{t-k}$, derive the matrix of partial correlations and draw up the CIG described by these.

Our aim is that CIG will assist us to select a structural VAR(p) for the series. However, the CIG only indicates undirected (conditional independence) links between variables (vertices). There may be several structural models (with direct dependence) which are consistent with the undirected graph and the important step is to determine these possibilities. We use the theory relating directed and undirected graphs, together with arrow of time and economic insight for this purpose.

In section 6.1 we examine the CIG and the DAG structure of the VAR(p) model, both in the canonical and in the structural form. In section 6.2 we shall show an illustrative example of how a DAG structure develops into a CIG model by moralisation. In section 6.3 we shall see how a structural VAR(p) model can be obtained via graphical modelling. Then we shall apply, in section 6.4, these techniques to a real dataset investigating the existence of the lending channel of the monetary transmission mechanism in the Italian economy. At the end, section 6.6, we examine the sample properties for multivariate time series models.

6.1 DAG and CIG structure of the VAR(p) models

In this section we consider the graphical structure implied by VAR(p) models. First consider a canonical VAR(1) model defined by the matrix equation

$$\begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} = \begin{pmatrix} \phi_{111} & \phi_{121} & \phi_{131} \\ \phi_{211} & \phi_{221} & \phi_{231} \\ \phi_{311} & \phi_{321} & \phi_{331} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \end{pmatrix} + \begin{pmatrix} e_t \\ f_t \\ g_t \end{pmatrix}. \tag{6.1.1}$$

The directed graph in figure 6.1 attempts to present this relationship. The dotted lines enclosing the two blocks of variables ($\boldsymbol{X}_t$, $\boldsymbol{Y}_t$, $\boldsymbol{X}_t$) and ($\boldsymbol{X}_{t-1}$, $\boldsymbol{Y}_{t-1}$, $\boldsymbol{X}_{t-1}$) represent the fact that the VAR(p) models block lagged relationship with a block error ($e_t,\ f_t,\ g_t$) which is independent of the explanatory variables. Similar links could be shown between

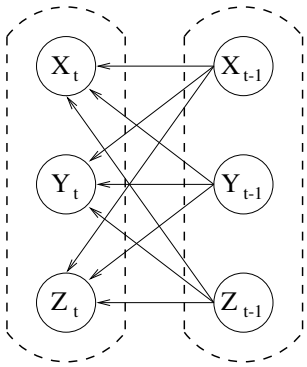


Figure 6.1: Graphical representation for a VAR(1) model.

time $t-1$ and time $t-2$ and so on, repeating the structure of the graph in figure 6.1 in a similar way to that in figure 5.1. The components of the block error are however, in general, correlated. The edges in the figure are included to represent the parameters ϕ_{ijk} (variable i regressed on variable j at lag k) associated with each edge.

In a canonical AR(1) model there are 9 coefficients, 3 residual error variances and 3 contemporaneous residual correlations. The residuals e_t, f_t, g_t are the errors of prediction of present observation from past values alone, using no contemporaneous dependence. In a directed graph representing a canonical VAR(p) model, each vertex at time t is linked to all the vertices at time $t-1,\ t-2, \ldots, t-p$.

A structural VAR model allows the possibility of contemporaneous dependence with a corresponding simplification of the correlation structure of the residuals which are assumed to be independent of each other. It will also be sparse in its use of explanatory coefficients.

Example 6.1.1 *Figure 6.2 shows the graphical model of a structural VAR(1).*

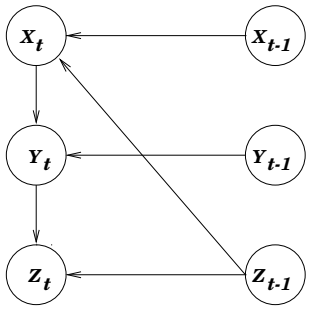


Figure 6.2: Graphical representation for a structural VAR(1) model.

This model can be represented by the matrix form

$$\begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ \phi_{210} & 0 & 0 \\ 0 & \phi_{321} & 0 \end{pmatrix} \begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} + \begin{pmatrix} \phi_{111} & 0 & \phi_{131} \\ 0 & \phi_{221} & 0 \\ 0 & \phi_{321} & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \end{pmatrix} + \begin{pmatrix} u_t \\ v_t \\ w_t \end{pmatrix}$$

which has a sparse residuals' coefficient matrix.

Moralising the DAG in figure 6.2 and considering just the edges at vertices at time t, we would obtain the CIG in figure 6.3. The labels on the edges are omitted; the graph is clear without them, □

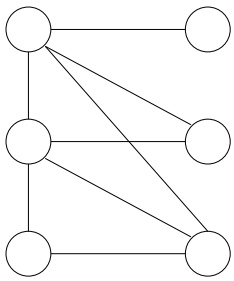


Figure 6.3: Moralizing the DAG in fig. 6.2.

Residuals from a canonical VAR contain relevant information, as proved by the following result.

Lemma 6.1.2 *Let $\boldsymbol{X}_t$ follow a multivariate model with canonical innovations $\boldsymbol{E}_t$. Then the partial correlations are equivalent:*

$$\text{Corr}\,(e_{it}, e_{jt} \mid \boldsymbol{E}_t \backslash \{e_{it}, e_{jt}\}) = \text{Corr}\,(x_{i,t}, x_{jt} \mid \boldsymbol{X}_t \backslash \{x_{i,t}, x_{jt}\}) \tag{6.1.2}$$

Proof:

$$\begin{aligned} \text{Corr}\,(e_{it}, e_{jt} \mid \boldsymbol{E}_t \backslash \{e_{it}, e_{jt}\}) &= \text{Corr}\,(e_{i,t}, e_{jt} \mid \boldsymbol{X}_t, \boldsymbol{X}_{t-1}, \ldots \backslash \{x_{i,t}, x_{jt}\}) \\ &= \text{Corr}\,(x_{i,t}, x_{jt} \mid \boldsymbol{X}_t, \boldsymbol{X}_{t-1}, \ldots \backslash \{x_{i,t}, x_{jt}\}). \quad (6.1.3) \end{aligned}$$

□

Corollary 6.1.3 *Let G be the CIG of $\boldsymbol{X}_t$, $\boldsymbol{X}_{t-1}, \ldots$ and G_t be the subgraph of G which contains only the contemporaneous variables $\boldsymbol{X}_t$ and their links. Then G_t is identical to the CIG of $\boldsymbol{E}_t$, the contemporaneous errors.* □

Then the CIG obtained by applying the inverse variance lemma to the correlation matrix of the residuals from a canonical VAR is equivalent to the CIG obtained by applying the same matrix manipulation to the original variables. In the seventh chapter we shall observe empirical evidence of this lemma.

The major value of the canonical VAR(p) model is its unique parametrisation. It is also encompassing of all structural VAR(p) model so that for practical applications in prediction it is simple to use, as stated before, without concern for a structural interpretation.

However, in many situations a structural VAR(p) model can be expected to be more parsimonious in parametrisation than the corresponding canonical VAR(p) besides having the advantage of interpretability. The difficulty is identification of the structural VAR(p) from historical data. Economic insight may suggest the form of model but could be incomplete or misleading. We wish to combine this with data based methods which guide us towards one, or possibly a small range of plausible structural models. These models may then be estimated, compared and improved using maximum likelihood methods. The data based methods we are now explaining are those using graphical modelling.

6.2 An illustrative example of a structural VAR(2)

The graph in figure 6.4 represents a structural VAR(2) model in the same manner as in the example of the AR(1). It is motivated by a simplistic model fitted to series which will be analysed in section 6.4. The coefficients are shown near the correspondent links.

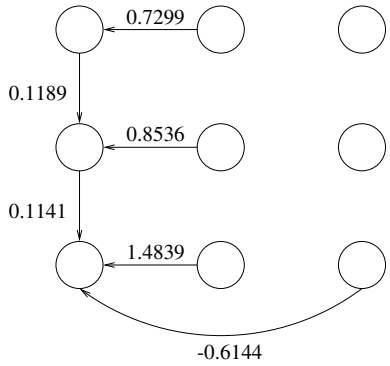


Figure 6.4: Hypothetical DAG.

The structural error covariance matrix is

$$\boldsymbol{D} = \begin{bmatrix} 1.144 & 0 & 0 \\ 0 & 0.1258 & 0 \\ 0 & 0 & 0.039 \end{bmatrix}. \tag{6.2.1}$$

This DAG structure can be extended to include variables at previous times. It is invariant in the way it links variables at a given time to those at previous times. Extending the DAG to include one extra time period, the corresponding CIG can be obtained by moralisation to be that shown in figure 6.5.

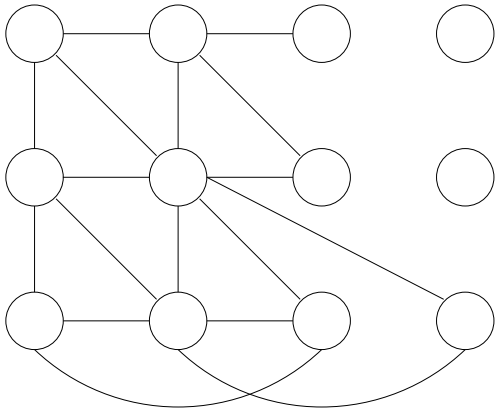


Figure 6.5: Theoretical moralisation of the DAG.

Moralisation has just been applied to structural raw links with variables shown in the diagram. Because there are links with previous variables not shown in the diagram, there will in fact be further links needed to complete this CIG. In the univariate case we could apply time reversal symmetry to complete these but this does not apply to multiple time series as $\mathrm{Cov}\,(x_t, x_{t-k})$ is not the same as $\mathrm{Cov}\,(x_{t-k}, y_t)$.

It might be possible to derive the CIG structure by creating a DAG covering a larger

time range and then collapsing this to the time range required (see Whittaker, 1990, pp. 394–401).

For this example we obtained the CIG shown in figure 6.6, with the virtual partial correlations entered. To obtain these values we calculated the autocovariance coefficients

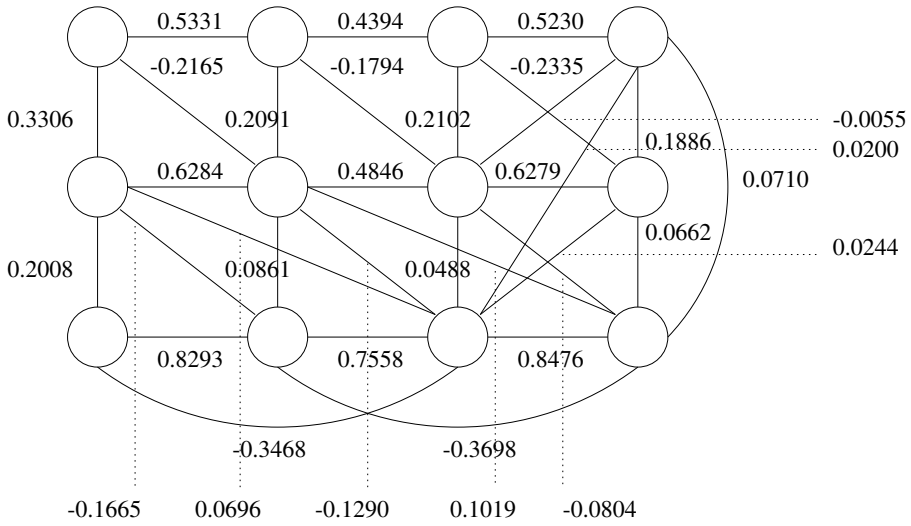


Figure 6.6: Hypothetical DAG moralised.

Γ_{xk} of the structural AR(2) model. To do this the model was written in state space form, i.e. as a 6-variable first order transition matrix:

$$\boldsymbol{Y}_t = \boldsymbol{T}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t$$

where

$$\boldsymbol{Y}_t^T = (X_t^T, X_{t-1}^T)$$

from which

$$\begin{aligned} \Gamma_y(0) &= \boldsymbol{V} + \boldsymbol{T}\boldsymbol{V}\boldsymbol{T}^T + \boldsymbol{T}^2\boldsymbol{V}\boldsymbol{T}^{2T} + \ldots \\ \Gamma_y(k) &= \boldsymbol{T}^k\Gamma_y(0), \qquad k = 0, 1, 2, \ldots \end{aligned}$$

From these the covariance matrix of $X_t, X_{t-1}, X_{t-2}, X_{t-3}$ was constructed and inverted to calculate the partial correlations.

The resulting CIG in figure 6.6 confirms the links already shown in figure 6.5 and completes the remaining links.

6.3 The CIG approach to structural VAR(p) model building

Graphical modelling can be used effectively to obtain sparse structural VARs. Once we have estimated a sample covariance matrix from the multivariate time series

$$\boldsymbol{X}_t, \boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-k},$$

we test for the significant entries in the partial correlation matrix and draw the correspondent CIG. We shall only take into considerations, for reasons explained before, links within present and between present and lagged variables. The results from Anderson (1971) about sampling properties of AR coefficients justify the application of the significance thresholds.

By "demoralisation", we consider the possible DAGs which might explain the CIG. In this way, links with lagged variables will affect possible structures between current variables.

In a structural VAR model with n current variables, there are n! possible recursive ways of linking them. So to find the most adequate subset structural VAR model requires a very large number of exploratory models if n is large. With graphical modelling we select a small number of possible DAGs whose suitability to the data generating process can be further assessed by maximum likelihood based identification methods.

6.4 An application: the lending channel of monetary transmission in Italy

We now consider a real example of three monetary time series for which two different economic theories lead to different structural models. We discuss the economic background and then apply graphical modelling to identify an empirical (statistical) structural model for these series.

6.4.1 The economic background

A monetary transmission mechanism is the means used by the economic policy maker to affect the *real economy* operating on monetary economics instruments. Until recent years, there was just one possible mechanism hypothesised in the economic literature.

For such a reason it is also called the *classic view* or the *classic channel*, but it is also called the *money channel*.

According to this we can distinguish only two different assets in the market: money and bonds; every other asset is a perfect substitute of one of them. In this scheme banks do not play any relevant role and are considered part of the household sector.
– In this case an increase of the repurchase agreement interest rate would force the bond interest rate up.
– The household sector, therefore, will not find convenient to transfer temporarily bonds to the central bank. As a matter of accounting, then, it must hold less money and more bonds. If there is not full and instantaneous adjustment of the prices there will be a loss of money in real terms for households; this will cause eventually an increase in real interest rates which in turn can have effects on investments and real economy.

Bernanke and Blinder in 1988 proposed a model for *aggregate demand* which allows the existence of another monetary transmission mechanism together with the existing *money channel*. It is called the *lending channel* or the *credit channel*. According to this theory bank loans and bonds are not perfect substitutes so that we can distinguish three different assets in the market: money, bonds and bank loans. Under these conditions monetary policy can work on either the bond interest rate or loan interest rate or both so that an impact on the latter can be independent from an effect on the former. Banks, in this scheme, play a relevant role and are not included in the household sector.
– In this case, an increase of the repurchase agreement interest rate will induce an expansion of banks' reserves.
– As shown by the monetary multiplier, such expansion will reduce the quantity of loans. If money and bonds are close substitutes there will be a minimal impact on bonds interest rate. Nevertheless the cut on loan supply will push up their cost with an influence on the real economy. In this case we have a weak money channel but a strong lending channel. A similar example can be found in Kashyap and Stein (1993).

In Bernanke and Blinder's model there are three necessary conditions for the existence of the lending channel:

1. from the firms point of view intermediated loans and open market bonds must not be perfect substitutes;

2. the central bank must be able, by changing the amount of reserves of the banking

system, to affect the supply of intermediated loans;

3. there must be an imperfect price adjustment to monetary policy shocks.

6.4.2 A previous analysis

There is a clear evidence that the Italian economy matches at least two of the above mentioned conditions and hence provides a suitable environment to verify the existence of the lending channel. In fact as Buttiglione and Ferri (1994) pointed out:

1. Italian firms, as their balance sheets show, are funded far more by banking credit than issued bonds or commercial paper, so that it is unlikely that they can unlikely be seen as perfect substitutes. There isn't any commercial paper market indeed;

2. during the eighties Italian banks reduced the amount of securities in their portfolios and now the adjustment is completed. Hence they can't neutralise a shock on reserves by asset management anymore.
   Moreover the lack of a secondary market for *deposit certificates* has prevented banks to use liability management in response to monetary restrictions.

The third condition, concerning the speed of price adjustment, although central for any monetary economics theory, is normally less apparent and more difficult to assess.

Recently Bagliano and Favero (forthcoming) carried out an empirical analysis to test whether the lending channel has worked in Italy. They estimated two different VAR models.

The first, of order two, has to investigate the transmission from the monetary policy impulse to the government bond and loan interest rate and hence to their difference. A widening of the latter implies the existence of the credit channel (see Bernanke and Blinder, 1988). It involves three variables: 1) the repurchase agreement interest rate (a), whose innovations may be viewed as monetary policy innovations; 2) the average interest rate on government bonds with residual life longer than one year (b); 3) the average interest rate on bank loans (c).

With the second VAR system, of order five, they assessed the impulses of monetary policy to real economy. It includes four different variables: 1) the difference between

bank loan and government bond; 2) the loan interest rate; 3) the industrial production; 4) the inflation.

6.4.3 The graphical modelling approach

We partially used Bagliano and Favero's framework to apply our VAR model identification strategy, investigating the relationships among the variables of the first VAR system they estimated, to verify if there is a direct causal effect from a monetary economics impulse (the repurchase agreement interest rate) to the loan interest rate. To pursue our analysis we used the same monthly time series taken from the same sources (Bank of Italy) over the period January 1986–December 1993.

Our objective is to model the relationships among the variables a, b and c including lagged variables up to lag 2 or higher, where the variables represent data vectors e.g. $a_t \equiv (a_3, \ldots, a_{96}), a_{t-1} \equiv (a_2, \ldots, a_{95}), a_{t-2} \equiv (a_1, \ldots, a_{94})$. A directed link connecting the repurchase agreement interest rate (lagged or current) to the loan interest rate would support the existence of the lending channel.

In our analysis we start by considering the relationship among current and lagged values of the REPO interest rate a_t, a_{t-1}, a_{t-2} of the bonds interest rate b_t, b_{t-1}, b_{t-2} anf of the loan interest rate c_t, c_{t-1} and c_{t-2}. Let's compute for them the sample correlation matrix $\boldsymbol{V}$

a_t	1.00								
a_{t-1}	0.78	1.00							
a_{t-2}	0.60	0.80	1.00						
b_t	0.64	0.64	0.57	1.00					
b_{t-1}	0.55	0.66	0.66	0.94	1.00				
b_{t-2}	0.45	0.56	0.67	0.82	0.93	1.00			
c_t	0.59	0.74	0.80	0.65	0.74	0.75	1.00		
c_{t-1}	0.42	0.61	0.75	0.52	0.66	0.74	0.96	1.00	
c_{t-2}	0.26	0.44	0.62	0.38	0.52	0.66	0.86	0.96	1.00
	a_t	a_{t-1}	a_{t-2}	b_t	b_{t-1}	b_{t-2}	c_t	c_{t-1}	c_{t-2}

The inverse of this matrix, V^{-1}, has diagonal elements whose value is $\frac{1}{1-r^2}$, where r is the multiple correlation coefficient between that variable and the rest (see Whittaker, 1990 pp. 3–6).

a_t	3.82								
a_{t-1}	-2.12	5.84							
a_{t-2}	-0.07	-2.52	4.77						
b_t	-2.99	2.14	-0.13	15.66					
b_{t-1}	3.95	-4.73	1.88	-16.84	32.91				
b_{t-2}	-1.27	2.71	-2.08	2.14	-15.79	15.02			
c_t	-3.28	-3.79	2.64	-11.28	3.59	6.04	61.5		
c_{t-1}	2.39	3.68	-8.36	19.78	-16.13	-4.02	-87.03	161.39	
c_{t-2}	0.41	-0.88	4.36	-7.93	11.85	-3.33	30.36	-73.44	41.99
	a_t	a_{t-1}	a_{t-2}	b_t	b_{t-1}	b_{t-2}	c_t	c_{t-1}	c_{t-2}

In this case, among variable at time t, the best explained is c_t for which $r^2 = \frac{61.5-1}{61.5} = 0.98$ and a_t is the least predictable with $r^2 = \frac{3.82-1}{3.82} = 0.74$, consistently with the economic theory.

As before, from V^{-1} we obtain the matrix W of the partial correlation coefficients between the corresponding pair of variables given the remaining ones.

a_t	1.00								
a_{t-1}	0.45	1.00							
a_{t-2}	0.02	0.48	1.00						
b_t	0.39	-0.22	0.01	1.00					
b_{t-1}	-0.35	0.34	-0.15	0.74	1.00				
b_{t-2}	0.17	-0.29	0.25	-0.14	0.71	1.00			
c_t	0.21	0.20	-0.15	0.36	-0.08	-0.20	1.00		
c_{t-1}	-0.10	-0.12	0.30	-0.39	0.22	0.08	0.88	1.00	
c_{t-2}	-0.03	0.06	-0.31	0.31	-0.32	0.13	-0.60	-0.89	1.00
	a_t	a_{t-1}	a_{t-2}	b_t	b_{t-1}	b_{t-2}	c_t	c_{t-1}	c_{t-2}

Now we can compute approximate critical values test statistic using the formula, $\frac{t^2}{t^2+n-k}$, where n is the series length, k the number of parameters and t the t-value, to investigate if these coefficients are different from zero. We used two threshold values: 0.26 (0.99 confidence interval) and 0.20 (0.95 confidence interval) distinguishing, in this way, between strong dependence ($\bullet$), weak dependence ($\circ$) and conditional independence ($\cdot$). We then obtain the following matrix

a_{t-1}	$\bullet$						
a_{t-2}	$\cdot$						
b_t	$\bullet$	$\circ$	$\cdot$				
b_{t-1}	$\bullet$			$\bullet$			
b_{t-2}	$\cdot$			$\cdot$			
c_t	$\circ$	$\circ$	$\cdot$	$\bullet$	$\cdot$	$\cdot$	
c_{t-1}	$\cdot$			$\bullet$			$\bullet$
c_{t-2}	$\cdot$			$\bullet$			$\bullet$
	a_t	a_{t-1}	a_{t-2}	b_t	b_{t-1}	b_{t-2}	c_t

where the test is performed just for the edges linking to variables at time t as, according to what we said before, the test is appropriate only in this case.

As before we represent this matrix using a conditional independence graph (see figure 6.7).

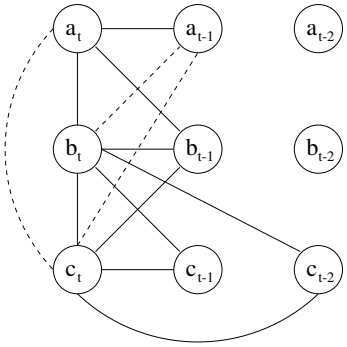


Figure 6.7: Graphical representation of the independence matrix.

where a_{t-2} and b_{t-2} do not have any direct effect on variables at time t. Then only the links between c_{t-2} and c_t and between c_{t-2} and b_t justify the use of the second order in the specification of the VAR model.

Our objective is to hypothesise structural models, which we represent by directed graphs, which are consistent with this undirected graph.

Adding the arrow of time where it is possible and eliminating vertices a_{t-2} and b_{t-2} we obtain the mixed graph, with both directed and undirected edges, in figure 6.8.

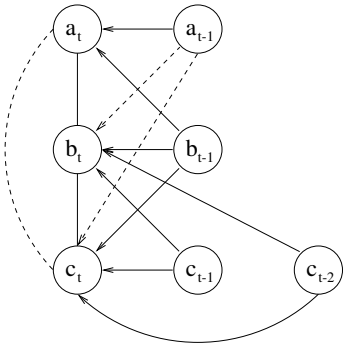


Figure 6.8: Mixed graph.

We must be aware that some of the links in this graph may be due to moralisation.

Hence what we have still to explain is the structure between current variables and which links are moral ones. Links between past and current variables have already a direction.

They do not matter in terms of moralisation of the links between current variables as current variables cannot be parents of lagged variables. The opposite is untrue, i.e. different structures for the current variables have relevant consequences for demoralisation of links between past and current variables.

The subgraph considering only the current variables and the links connecting them is complete (see figure 6.9) and because of that there are several (9) directed alternative

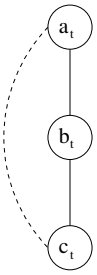


Figure 6.9: Subgraph for current variables.

models which can explain it and which are shown in figure 6.10.

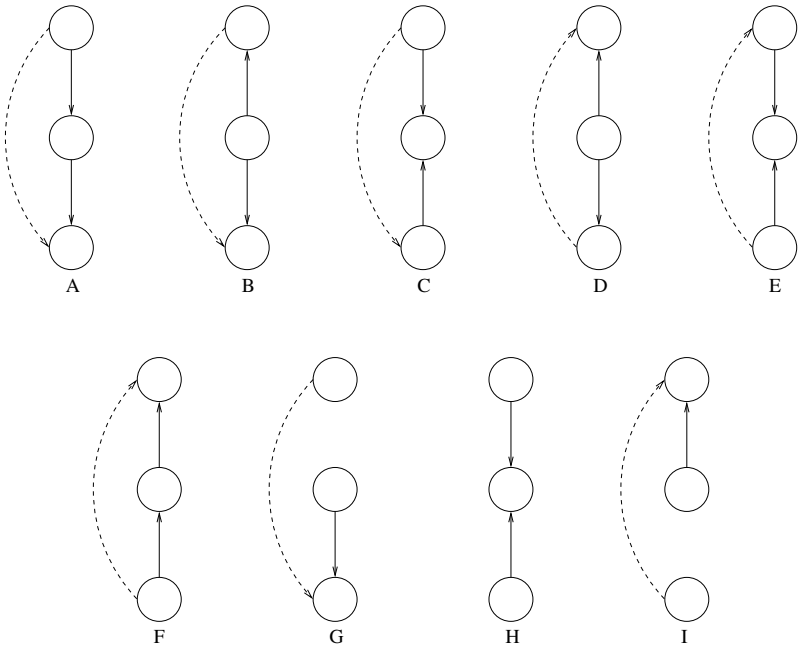


Figure 6.10: Alternative directed graphs.

The next step is checking which DAGs are compatible with the original CIG considering, one by one, the hypothesised sub-DAGs for the current variables. In finding the compatible models which imply different moral links, three considerations may be handy.

They are a consequence of the context in which we are working, i.e. time series. As the graph in figure 6.8 shows, when we apply Graphical Modelling techniques to time series, we may be able to give an *a priori* direction to edges whose existence is just due to moralisation and that, in fact, will not appear in the DAG. In the standard context, where no lagged are present together with the correspondent current ones, once we have given a direction to a link, we have already decided that it is real (not moral). This is not the case when GM is applied to time series where the a priori direction can help us in deciding if a link is moral or real.

The three considerations are:

first consideration. *If a vertex has just one outgoing arrow, that is not a moral link.* This is because a moral edge links two parents, that is two vertices which both have at least another (outgoing edge).

second consideration. *If a vertex has more than one outgoing edge, at least one of them is not moral; it is, of course, possible that some of them are moral.* The reason again is that moral edges link parents. Then at least one of the outgoing links is due to parenthood and hence it is real.

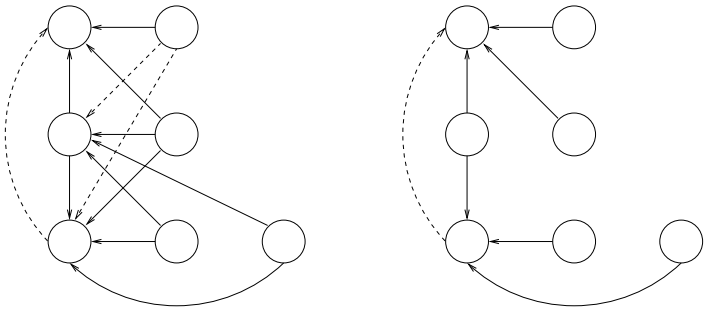
third consideration. *All the incoming oriented edges of a vertex with no outgoing edges are not moral.* If a vertex has no outgoing edges, it is not a parent of any edge.

Let us now apply these considerations in order to find feasible DAGs which are consistent with the original CIG (figure 6.8).

There are no compatible DAGs hypothesising the subgraph A. For the third consideration, in fact, the links connecting a_t, c_{t-1} and c_{t-2} are real. Nevertheless that would imply moral links between a_t and c_{t-1} and between a_t and c_{t-2} which are not present in the original CIG. For similar reasons there are no compatible DAGs including subgraphs B, C, E, G and H.

There are consistent DAGs containing the subgraph D. The most explicative of them is D1 (figure 6.11). It has 13 parameters (links), containing therefore the same edge set than the original CIG. The less explicative DAG containing D is D2 (figure 6.11). It has 7 parameters and none of them can be removed without being inconsistent with the original CIG. Hence all the DAGs containing D can be obtained by removing links from D1 while still containing D2. This makes easy to calculate the number of possible DAGs

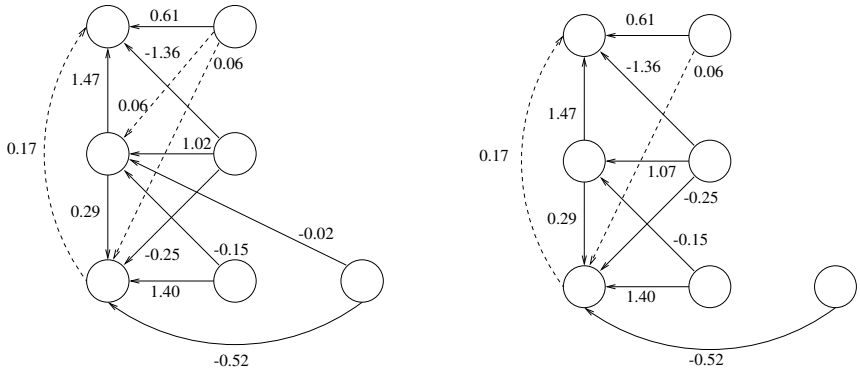
containing D, it is $2^{13-7} = 2^6 = 64$.



D1 D2

Figure 6.11: DAGs containing subgraph D.

We know already that D1 is the most explicative model, nevertheless we have still to assess which is the most parsimonious. In order to do that we can use likelihood based methods as described in section 5.4. In the specific case, we shall select as most parsimonious the model which will minimise the Schwartz information criterion. Instead of computing such an algorithm for all the 64 alternative models we shall take out, one by one, the less significant links from D1, which encompasses all the other 63 models, until, taking out a link, we increase the SIC. In figure 6.12 we show the most parsimonious model selected (D3) together with D1, both with the links estimated by OLS.

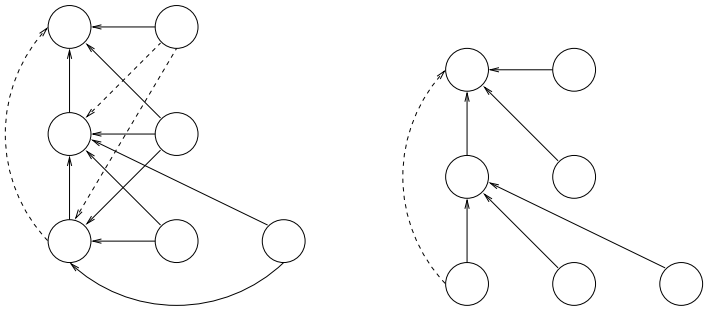


D1 D3

Figure 6.12: Most explicative (D1) and most parsimonious (D3) model containing subgraph D.

There are 64 DAGs consistent with the original CIG which contain the subgraph F. Using arguments similar to those for DAGs containing D, they are obtained as those lying between the most explicative model (F1), which has 13 parameters, and the explicative

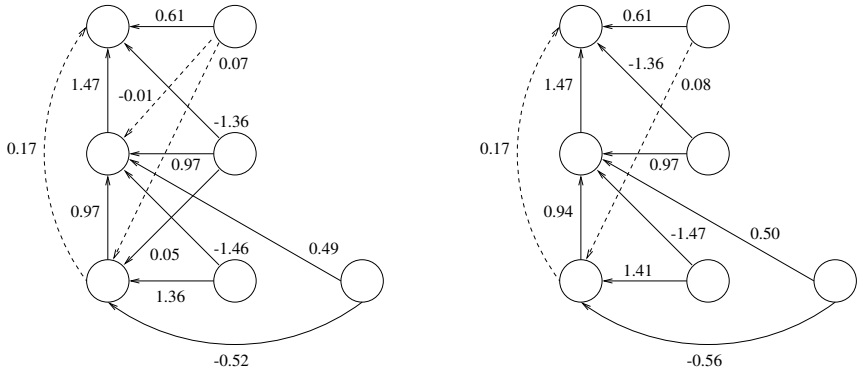
one (F2), which has 7 parameters. They are shown in figure 6.13.



F1 F2

Figure 6.13: DAGs containing subgraph F.

By subset regression, as for the previous case, starting from F1 we used SIC to select the most parsimonious model (F3). F1 and F3 are shown with link estimates in figure 6.14.



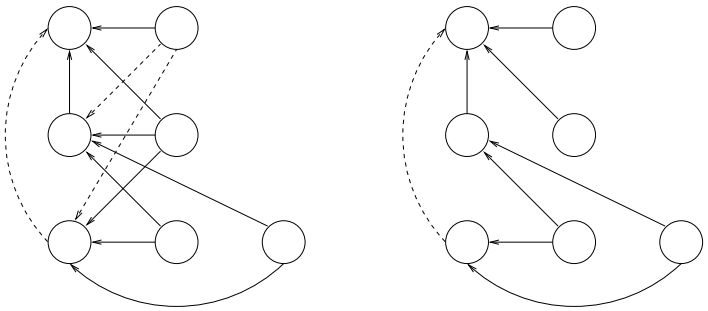
F1 F3

Figure 6.14: Most explicative (F1) and most parsimonious (F3) model containing subgraph F.

Repeating again the same arguments used we identified $2^4 = 16$ different alternative models containing subgraph I and consistent with the original CIG. They are obtained, as before, as those lying between the the most explicative model (I1) and the least explicative model (I2), both shown in figure 6.15.

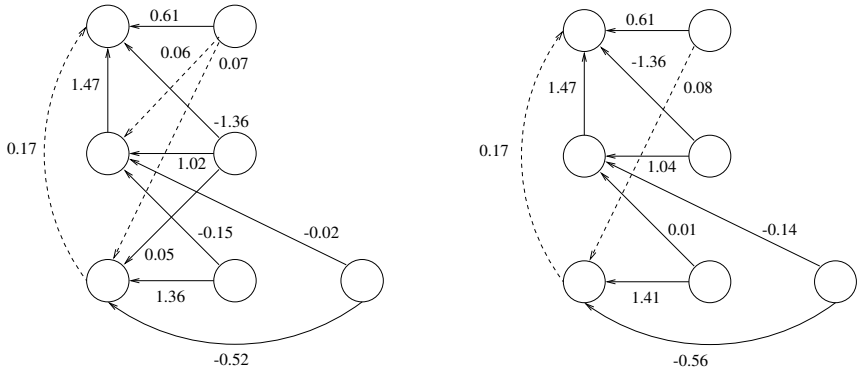
The most parsimonious model (I3) and I1 with parameter estimates are shown in figure 6.16.

Table 6.1 compares the different models in terms of likelihood based methods. A first



I1 I2

Figure 6.15: DAGs containing subgraph I.



I1 I3

Figure 6.16: Most explicative (I1) and most parsimonious (I3) model containing subgraph I.

consideration is that with the graphical modelling approach we have reduced the number of possible models from $2^{21} \times 6 = 12,582,912$ to $2^6 + 2^6 + 2^4 = 144$ with a reduction of about 87,000 to 1. A second consideration is that just testing the inverse variance matrix we obtain adequate models with a much smaller number of parameters, leaving to subset regression just a task of refinement. From the observation of the values in the

Table 6.1: Comparisons of the different models.

model	par.	lik	AIC	SIC	HIC
saturated	21	7.52	749.00	802.41	897.81
D1	13	7.73	752.86	785.89	844.95
D2	7	10.31	983.31	1001.1	1032.9
D3	11	7.78	753.04	781.01	830.99
F1	13	7.73	752.83	785.89	844.95
F2	7	13.19	1253.6	1271.4	1303.2
F3	11	7.76	751.50	779.48	829.45
I1	12	8.06	781.76	812.28	866.80
I2	8	9.93	949.43	969.77	1006.1
I3	10	8.12	782.99	808.42	853.86

table, it would appear that models incorporating subgraphs D and F are close, giving a sensibly better representation of the data generating process than models incorporating I. Following the considerations on the criteria in chapter 5 we attach greatest importance to models with lowest SIC. Such a criterion appears more balanced than the AIC, which tends to prefer overparametrised models, and the HIC, which is extremely penalising for extra parameters. This focuses particularly on D1, D3, F1 and F3, all with values close to 790 and less than that of the saturated model. Further comparisons can be made observing forecasting performances.

In each case, anyway, the weak link between a_{t-1} and c_t would give support, even if feeble, for the existence of the lending channel. The evidence of causal relation from c_t to a_t, in disagreement with the economic theory, could suggest the inadequacy of the monthly frequency used for the investigation.

To simplify our analysis we may, at some stage, use economic insight to direct our search for possible alternative models.

6.5 Further issues

As a generalisation of its use to identify sparse structures for time series models, graphical modelling can be used to identify also the order of the system. As an illustration we generated the undirected graph using variables up to lag 3 from the Italian interest rates data, it is shown in figure 6.17.

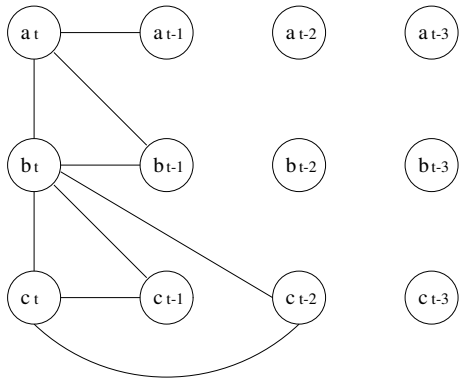


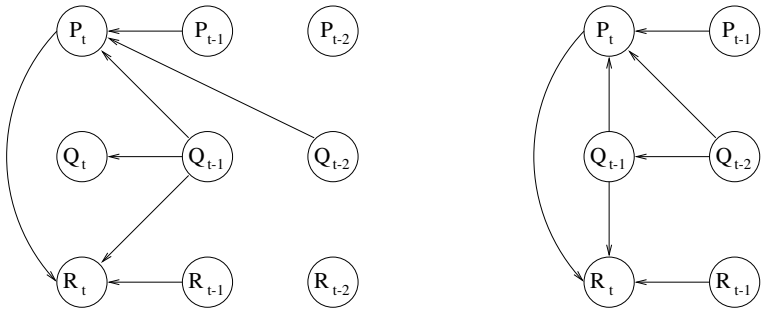
Figure 6.17: Graphical representation for a higher order independence matrix.

As we can see, none of the variables at time t-3 is adjacent to a_t, b_t and c_t. This is a further confirmation that a second order specification is correct.

The graphical modelling approach also allows us investigate of the effects of possible relative time shift of the variables which may reduce the order of the canonical form. This has the potential for reducing the estimator bias (see Abadir *et al.*, forthcoming). An example is shown in figure 6.18. Observing the DAG 6.18.1 we notice that as variable Q_t depends just on the its lagged variable Q_{t-1} we can shift $\{Q_t\}$ without without loosing any understanding of the mechanism. We do not have any advantage in the structural form as the number of parameters is the same both in the shifted and the original sparse model. Nevertheless we have an advantage if we decide to estimate the model in its canonical form, where we would have 9 links instead of 18.

6.6 Extended theory of sampling properties

Our objective in this section is to establish the large sampling properties of the estimates of the parameters in the structural autoregressive models when estimated by single equation OLS. This justifies the significance thresholds which we have used in the graphical modelling procedures of the previous sections in this chapter. The proof is an extension



6.18.1 6.18.2

Figure 6.18: Shifting variables in a structural VAR.

of the theorem given by Anderson (1971, chapter 5) for the derivation of the same result for canonical autoregressive models.

Theorem 6.6.1 *Let $\boldsymbol{X}_t$ follow the structural autoregressive model*

$$\boldsymbol{\Phi}_0 \boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \ldots + \boldsymbol{\Phi}_p \boldsymbol{X}_{t-p} + \boldsymbol{E}_t \tag{6.6.1}$$

where $\boldsymbol{E}_t$ is a vector of independent Gaussian series and let $\hat{\boldsymbol{\Phi}}_{ijk}$ be the estimate of $\boldsymbol{\Phi}_{ijk}$ obtained by OLS applied to the single equation

$$\boldsymbol{X}_{t,i} = -\sum \boldsymbol{\Phi}_{ij0} \boldsymbol{X}_{t,j} + \sum_k \sum_j \boldsymbol{\Phi}_{ijk} \boldsymbol{X}_{t-k,j} + \boldsymbol{E}_{t,i}. \tag{6.6.2}$$

Then for $t = 1 + p, \ldots, n$ and for each fixed i, the vector of estimates $\{\hat{\boldsymbol{\Phi}}_{ijk}\}$ has the approximate large sample distribution

$$\{\hat{\boldsymbol{\Phi}}_{ijk}\} \sim MVN\left(\{\boldsymbol{\Phi}_{ijk}\},\ \frac{1}{n}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \sigma_e^2\right) \tag{6.6.3}$$

where $\boldsymbol{X}$ is the design matrix of the regression 6.6.2 and $\sigma_e^2 = Var(\boldsymbol{E}_{t,i})$.

Proof:

Lemma 6.6.2 *If $\boldsymbol{X}_t$ is generated by the structural model 6.6.1, the error $\boldsymbol{E}_{t,i}$ in the single equation regression 6.6.2 is independent of the regressors $\boldsymbol{X}_{t,j}$ and $\boldsymbol{X}_{t-k,j}$ in that equation.*

Proof: We have previously shown how this can be transformed into a canonical autoregression. The assumption, in the canonical VAR, that the errors $\boldsymbol{E}_t$ form an independent sequence implies that they are also independent of past regressors $\boldsymbol{X}_{t-k}$. In the structural VAR we assume that the contemporaneous errors $\boldsymbol{E}_{t,i}$ are independent of past

errors $\boldsymbol{E}_{t-k}$ and independent of each other. Because of the causal DAG structure, this similarly implies that the errors $\boldsymbol{E}_{t,i}$ are also independent of the contemporaneous regressors $\boldsymbol{X}_{t,j}$, as well as of the past values $\boldsymbol{X}_{t-k,j}$, in the single equation regression for $x_{t,i}$. □

Lemma 6.6.3

$$plim\frac{1}{n}\sum \boldsymbol{E}_{t,j}\boldsymbol{X}_{t-k,j} = 0 \tag{6.6.4}$$

for all the regressors (past and contemporaneous) $\boldsymbol{X}_{t-k,j}$ in the single equation regression 6.6.2 for $\boldsymbol{X}_{t,i}$.

Proof: This depends only on the stationarity of $\boldsymbol{X}_t$ and the independence of $\boldsymbol{E}_{t,i}$ and $\boldsymbol{X}_{t-k,j}$, which immediately gives the expectation

$$\frac{1}{n}\sum \boldsymbol{E}_{t,i}\boldsymbol{X}_{t-k,j} \qquad\qquad ,k = 1,\ldots,p \tag{6.6.5}$$

to be zero and the variance to be

$$\frac{1}{n}\sigma^2_{e_i}\sigma^2_{x_j},$$

following paragraph 5.5.3 in Anderson. This tends to zero as $n \to \infty$. □

Lemma 6.6.4 *(Consistency)*

$$plim\hat{\boldsymbol{\Phi}}_{ijk} = \boldsymbol{\Phi}_{ijk}$$

Proof: The least squares equations for model 6.6.2, on dividing by n are

$$\frac{1}{n}\sum_t \boldsymbol{X}_{t,i}\boldsymbol{X}_{t-l,m} = -\sum_i \hat{\boldsymbol{\Phi}}_{ij0}\frac{1}{n}\sum_t \boldsymbol{X}_{t,j}\boldsymbol{X}_{t-l,m} + \sum_k\sum_j \hat{\boldsymbol{\Phi}}_{ijk}\frac{1}{n}\sum_t \boldsymbol{X}_{t-k,j}\boldsymbol{X}_{t-l,m},$$

which are obtained by setting

$$\sum_t \hat{\boldsymbol{E}}_{t,i}\boldsymbol{X}_{t-l,m} = 0$$

for all the regressors $\boldsymbol{X}_{t-l,m}$ in the equation 6.6.2. Substituting for $\boldsymbol{X}_{t,i}$ from 6.6.2 gives

$$\begin{aligned}\frac{1}{n}\sum \boldsymbol{E}_{t,i}\boldsymbol{X}_{t-l,m} = \; &- \sum_j \left(\hat{\boldsymbol{\Phi}}_{ij0} - \boldsymbol{\Phi}_{ij0}\right)\frac{1}{n}\boldsymbol{X}_{t,j}\boldsymbol{X}_{t-l,m} + \\ &+ \sum_k\sum_j \left(\hat{\boldsymbol{\Phi}}_{ijk} - \boldsymbol{\Phi}_{ijk}\right)\frac{1}{n}\sum_t \boldsymbol{X}_{t-k,j}\boldsymbol{X}_{t-l,m}.\end{aligned} \tag{6.6.6}$$

From 6.6.3, the LHS$\longrightarrow$ 0 so RHS$\longrightarrow$ 0. Provided only that the matrix of the resulting equations in $\boldsymbol{\delta}_{ijk} = \left(\hat{\boldsymbol{\Phi}}_{ijk} - \boldsymbol{\Phi}_{ijk}\right)$ has a non-singular limit (see below), the result is that the differences $\longrightarrow$ 0 so $\hat{\boldsymbol{\Phi}}_{ijk} \longrightarrow \boldsymbol{\Phi}_{ijk}$. This follows Anderson (par. 5.5.3). □

Lemma 6.6.5 *Let* $\boldsymbol{M}$ *be the matrix with entries*

$$\text{Cov}\left(\boldsymbol{X}_{t-l_1,m_1}\boldsymbol{X}_{t-l_2,m_2}\right)$$

where $\boldsymbol{X}_{t-l_1,m_1}$ *and* $\boldsymbol{X}_{t-l_2,m_2}$ *are any pair of regressors in 6.6.2. Then the vector of quantities*

$$\boldsymbol{Z} = \frac{1}{\sqrt{n}}\sum_i \boldsymbol{E}_{t,i}\boldsymbol{X}_{t-l,m}$$

has the asymptotic distribution

$$\boldsymbol{Z} \sim MVN(0, \boldsymbol{M}\sigma_e^2).$$

Proof: Transforming 6.6.1 into a canonical AR allows $\boldsymbol{X}_t$ to be approximated by a large order moving average $\boldsymbol{X}_t^{(q)}$ so that a *central limit theory* may be applied to the sums

$$\frac{1}{\sqrt{n}}\sum \boldsymbol{E}_{t,i}\boldsymbol{X}_{t-k,j} \qquad\qquad ,k = 1,\ldots,p$$

following 5.5.4 in Anderson, where n is the series length. The only requirement is that the model for $\boldsymbol{X}_t$ should be stationary and that $\boldsymbol{E}_{t,i}$ is independent of $\boldsymbol{X}_{t-k,j}$ as demonstrated in lemma 6.6.2. The covariance between the terms

$$\frac{1}{\sqrt{n}}\sum_t \boldsymbol{E}_{t,i}\boldsymbol{X}_{t-l_1,m_1} \quad \text{and} \quad \frac{1}{\sqrt{n}}\sum_t \boldsymbol{E}_{t,i}\boldsymbol{X}_{t-l_2,m_2},$$

for two different regressors also reduces to

$$\boldsymbol{\sigma}_e^2\text{Cov}\left(\boldsymbol{X}_{t-l_1,m_1}, \boldsymbol{X}_{t-l_2,m_2}\right).$$

Thus $\boldsymbol{M}\sigma_e^2$ is the asymptotic covariance matrix in the central limit theorem. □

Lemma 6.6.6 *Let the random vectors* δ *have elements* $(\hat{\boldsymbol{\Phi}}_{ijk} - \boldsymbol{\Phi}_{ijk})$. *Then asymptotically*

$$\sqrt{n}\delta \sim MVN(0, \boldsymbol{M}^{-1}\sigma_e^2).$$

Proof: Rewrite equation 6.6.6 as

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum \boldsymbol{E}_{t,i}\boldsymbol{X}_{t-l,m} = \; & - \sum_j \sqrt{n}\boldsymbol{\delta}_{ij0}\frac{1}{n}\boldsymbol{X}_{t,j}\boldsymbol{X}_{t-l,m} + \\ & + \sum_k\sum_j \sqrt{n}\boldsymbol{\delta}_{ijk}\frac{1}{n}\sum_t \boldsymbol{X}_{t-k,j}\boldsymbol{X}_{t-l,m}. \qquad (6.6.7)\end{aligned}$$

The central limit theorem applies to the LHS.

The matrix on the RHS of the equation 6.6.7 has the form $\boldsymbol{X}^T\boldsymbol{X}$, with elements

$$\frac{1}{n}\sum \boldsymbol{X}_{t-j,k}\boldsymbol{X}_{t-1,m}. \tag{6.6.8}$$

By the ergodicity of $\boldsymbol{X}_t$ this also converge to $\boldsymbol{M}$ derived in lemma 6.6.5. In large samples therefore 6.6.7 is approximated by

$$\boldsymbol{Z} = \boldsymbol{M}\sqrt{n}\boldsymbol{\delta}$$

so that from lemma 6.6.5

$$\boldsymbol{Z} \sim \text{MVN}(0, \boldsymbol{M}\boldsymbol{\sigma}_e^2) \Longrightarrow \sqrt{n}\boldsymbol{\delta} = \boldsymbol{M}^{-1}\boldsymbol{Z} \sim \text{MVN}(0, \boldsymbol{M}^{-1}\boldsymbol{\sigma}_e^2).$$

□

The proof of theorem 6.6.1 follows almost directly from the last lemma. By using the approximation for $\boldsymbol{M}$, in large samples

$$\sqrt{n}(\hat{\boldsymbol{\Phi}}_{ijk} - \boldsymbol{\Phi}_{ijk}) \sim \text{MVN}(0, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma_e^2) \Rightarrow \hat{\boldsymbol{\Phi}}_{ijk} \sim \text{MVN}(\boldsymbol{\Phi}_{ijk}, \frac{1}{n}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma_e^2). \tag{6.6.9}$$

□

Chapter 7

Graphical Modelling Approach to Multivariate ARMA Models

In this chapter we present our "first contact" in extending the graphical modelling approach to multivariate ARMA models. We consider how we can construct a CIG for the relationship between current and lagged values of the series when moving average terms are present. A structural model form is then identified by extending the method used in the previous chapter for multivariate AR models.

7.1 Structural VARMA models

In chapter 3 we introduced the canonical VARMA model and illustrated them by fitting a VARMA(1,1) to a series of term rates:

$$\boldsymbol{X}_t = \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{E}_t - \boldsymbol{\Theta}\boldsymbol{E}_{t-1}. \tag{7.1.1}$$

Our objective in this chapter is to use GM to develop models similar to this which allow contemporaneous dependence in $\boldsymbol{X}_t$ in the same way as structural models, i.e.

$$\boldsymbol{\Phi}_0\boldsymbol{X}_t = \boldsymbol{\Phi}_1\boldsymbol{X}_{t-1} + \boldsymbol{A}_t - \boldsymbol{\Theta}_1\boldsymbol{A}_{t-1}. \tag{7.1.2}$$

In this structural VARMA we replace $\boldsymbol{E}_t$ which are canonical residuals with contemporaneous correlation, by $\boldsymbol{A}_t$ which are structural residuals contemporaneously uncorrelated or orthogonal. The parameters $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Theta}_1$ in equation 7.1.2 are not the same as $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ in 7.1.1. The relationship between the two models is given by multiplying 7.1.1 by $\boldsymbol{\Phi}_0$, then

$$\boldsymbol{\Phi}_1 \;=\; \boldsymbol{\Phi}_0\boldsymbol{\Phi}$$

$$\begin{aligned} \boldsymbol{A}_t &= \boldsymbol{\Phi}_0 \boldsymbol{E}_t \\ \boldsymbol{\Phi}_0 \boldsymbol{\Theta} \boldsymbol{E}_{t-1} &= \boldsymbol{\Theta}_1 \boldsymbol{A}_{t-1} = \boldsymbol{\Theta}_1 \boldsymbol{\Phi}_0 \boldsymbol{E}_{t-1} \Longrightarrow \boldsymbol{\Phi}_0 \boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \boldsymbol{\Phi}_0. \end{aligned} \tag{7.1.3}$$

Our main interest is in determining $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Phi}_1$. We hope that these will be sparse and that they can be interpreted, in DAG form, as the causal dependence of present values $\boldsymbol{X}_t$ on other contemporaneous values of $\boldsymbol{X}_t$ and past values of $\boldsymbol{X}_{t-1}$. Our interest in $\boldsymbol{\Theta}_1$ is less because it is more difficult to interpret. Equation 7.1.3 shows that $\boldsymbol{E}_{t-1}$ and $\boldsymbol{A}_{t-1}$ are linearly related so that in the absence of sparse structure in $\boldsymbol{\Theta}$ or $\boldsymbol{\Theta}_1$ there is no statistical preference for using $\boldsymbol{E}_{t-1}$ or $\boldsymbol{A}_{t-1}$ in the equation. As we investigate the structural model we know $\boldsymbol{E}_{t-1}$ but do not know $\boldsymbol{A}_{t-1}$, because the definition of $\boldsymbol{A}_t$ depends on the choice of the structural coefficients $\boldsymbol{\Phi}_0$, which we have still to identify. We therefore propose an intermediate, or hybrid model of the form

$$\boldsymbol{\Phi}_0 \boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{A}_t - \tilde{\boldsymbol{\Theta}} \boldsymbol{E}_{t-1} \tag{7.1.4}$$

where $\tilde{\boldsymbol{\Theta}} = \boldsymbol{\Phi}_0 \boldsymbol{\Theta}$, which we can identify using GM methods. The variables which are used in the GM approach will then be the components of $\boldsymbol{X}_t$, $\boldsymbol{X}_{t-1}$ and $\boldsymbol{E}_{t-1}$. The use of the lagged values $\boldsymbol{X}_{t-1}$ of the series and $\boldsymbol{E}_{t-1}$ of estimated residuals was considered by Durbin (1960) as a means of estimation of a multivariate ARMA model by ordinary least squares. We now extend this idea to the multivariate context, but with the novel aim of applying GM to the same variables. This will lead, demoralisation, to the identification of the hybrid structural model 7.1.4.

7.2 Efficient structural VARMA model identification using graphical modelling

Durbin (1960) proposed a method for estimating a univariate ARMA model for x_t by first obtaining consistent estimates $\hat{a}_t$ of the residual series a_t, by fitting a high order autoregression, then carrying out an ordinary regression

$$x_t = \phi_1 x_{t-1}, \ldots, \phi_p x_{t-p} + \theta_1 \hat{a}_{t-1} + \ldots + \theta_q \hat{a}_{t-q} + e_t.$$

He demonstrates that this provided consistent though not fully efficient estimates of the parameters. This suggests that a natural extension of the GM approach to VARMA models would then be to include, as the set of variables to which the GM is applied, both the series variables $\boldsymbol{X}_t, \boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-p}$ and the lagged innovations $\boldsymbol{E}_{t-1}, \ldots, \boldsymbol{E}_{t-q}$.

In this chapter we apply this method to the hybrid VARMA(1,1) model in 7.1.4. Rather than using a high order AR model we use the estimates of $\boldsymbol{E}_t$ obtained from efficient maximum likelihood estimation of the canonical VARMA(1,1) model.

Our procedure then is to form the covariance matrix for $\boldsymbol{X}_t$, $\boldsymbol{X}_{t-1}$ and $\boldsymbol{E}_{t-1}$, then derive the partial correlations and draw up a CIG for these variables. We then consider the possible DAG interpretations in terms of the hybrid model 7.1.4.

This method follows Durbin's approach to ARMA models estimation described above and therefore also reflects the inefficiency of his methodology. The efficiency can be improved by reference to the maximum likelihood estimation described in chapter 3. We use, instead of the variables $\boldsymbol{X}_t$, $\boldsymbol{X}_{t-1}$ and $\boldsymbol{E}_{t-1}$, respectively the constructed variables $\boldsymbol{U}_t$, $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ defined by

$$\boldsymbol{V}_t = -\boldsymbol{E}_{\Phi,t} \tag{7.2.1}$$

$$\boldsymbol{W}_t = -\boldsymbol{E}_{\Theta,t} \tag{7.2.2}$$

$$\boldsymbol{U}_t = \boldsymbol{E}_t + \boldsymbol{\Phi}_0\boldsymbol{V}_t + \boldsymbol{\Theta}_0\boldsymbol{W}_t. \tag{7.2.3}$$

These are chosen because the regression of $\boldsymbol{E}_t$ upon $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ gives the fully efficient estimates of the parameter corrections $\delta\hat{\boldsymbol{\Phi}} = \hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0$ and $\delta\hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0$ as described in chapter 3 by the equation 3.6.24, where $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Theta}_0$ are current parameter values. Defining $\boldsymbol{U}_t$, by adding $\boldsymbol{\Phi}_0\boldsymbol{V}_t + \boldsymbol{\Theta}_0\boldsymbol{W}_t$ to $\boldsymbol{E}_t$, results in a regression of $\boldsymbol{U}_t$ upon $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ which gives fully efficient estimates of $\hat{\boldsymbol{\Phi}}$ and $\hat{\boldsymbol{\Theta}}$, rather than the parameter corrections $\delta\hat{\boldsymbol{\Phi}}$ and $\delta\hat{\boldsymbol{\Phi}}$, i.e.

$$\boldsymbol{U}_t = \boldsymbol{\Phi}\boldsymbol{V}_t + \boldsymbol{\Theta}\boldsymbol{W}_t + \boldsymbol{E}_t. \tag{7.2.4}$$

We shall also consider a hybrid form of this as an efficient means of identifying and estimating the hybrid form 7.1.4:

$$\boldsymbol{\Phi}_0\boldsymbol{U}_t = \boldsymbol{\Phi}_1\boldsymbol{V}_t + \tilde{\boldsymbol{\Theta}}\boldsymbol{W}_t + \boldsymbol{A}_t. \tag{7.2.5}$$

On the basis that replacing $\boldsymbol{X}_t$, $\boldsymbol{X}_{t-1}$ and $\boldsymbol{E}_{t-1}$ by $\boldsymbol{U}_t$, $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ replaces inefficient parameter estimation by efficient parameter estimation; we also use $\boldsymbol{U}_t$, $\boldsymbol{V}_t$, $\boldsymbol{W}_t$ for a more efficient construction of the CIG.

7.3 The application to the term structure of the U.S. dollar interest rate

In this section we present results of the approaches just described. The series, for which we estimated a VARMA(1,1) model in chapter 3 are the seven different terms to maturity of the U.S. dollar interest rate. We first use the construction which follows Durbin's approach and then the one with the improved efficiency. In order to obtain the CIG we therefore first applied the inverse variance lemma to the correlation matrix of current variables, lagged variables and lagged errors obtained from the canonical VARMA(1,1):

$$\boldsymbol{X}_t, \boldsymbol{X}_{t-1}, \boldsymbol{E}_{t-1}.$$

The partial correlation between current and lagged variables is, therefore, affected by the estimate of the error component. We follow this by the more efficient approach using instead

$$\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t.$$

In figure 7.1 we draw the CIG including only current and lagged variables. The significant links obtained with the first, less efficient, procedure are represented with solid lines. Using the second, more efficient, procedure we obtained all the significant links obtained with the first procedure plus some extra-links; these are represented with dotted lines.

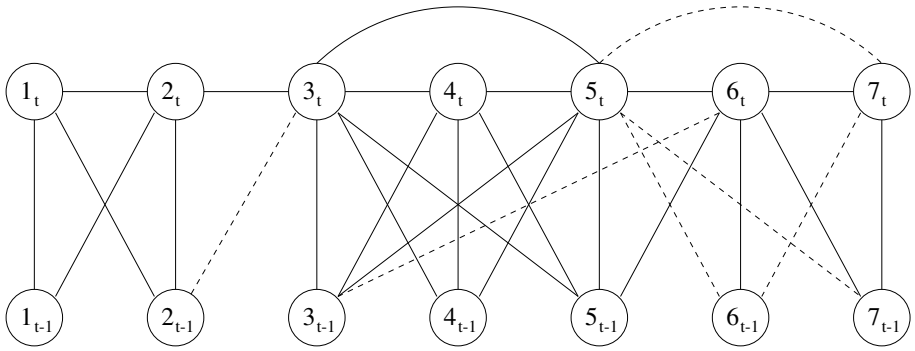


Figure 7.1: CIG deriving from a VARMA(1,1) model for the U.S. dollar interest rate.

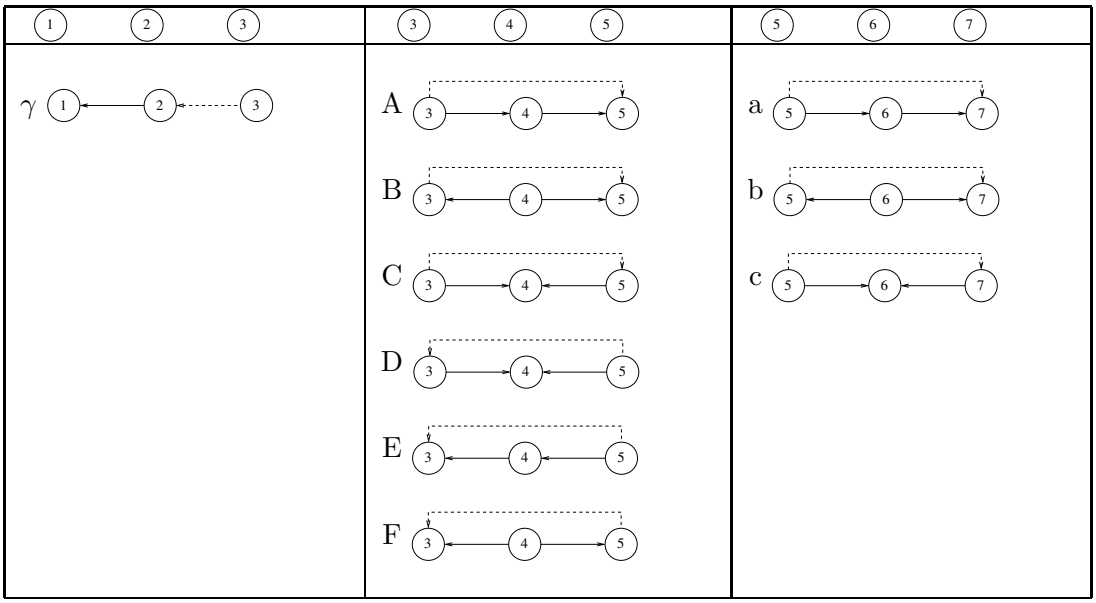
The subgraph of the current variables in the CIG obtained with the improved likelihood is exactly the same we examined in chapter 5, for the canonical residuals $\boldsymbol{E}_t$, confirming empirically, in this way, the results of lemma 6.1.2.

The efficient method shows that this must happen, because conditioned on $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$, the constructed variable $\boldsymbol{U}_t$ is equivalent to $\boldsymbol{E}_t$ from 7.2.3.

We now wish to use the CIG to obtain a DAG which identifies the non-zero coefficients of $\mathbf{\Phi}_0$ and $\mathbf{\Phi}_1$, in the structural VARMA(1,1) model. Similarly to what we did previously, we can use the arrow of time to give a direction to links between current and lagged variables but we have to find out the direction of links between current variables. We can use the analysis we did in chapter 4 and the presence of lagged variables will help us to restrict the number of possible models as, again, the DAG we hypothesise must be consistent with the original CIG in figure 6.1.2.

In particular, with reference to table 4.3 subgraphs α and β are ruled out because of the absence of a link between edges 2_t and 3_{t-1}. Also, subgraphs d, e and f are ruled out because of the absence of a link between 7_t and 5_{t-1}. The remaining possible subgraphs are represented in table 7.1.

Table 7.1: Possible directed subgraphs.



This reduces the number of possible models from 28 to 14; they are:
1)γAa; 2)γAc; 3)γBa; 4)γBc; 5)γCa; 6)γCc; 7)γDa; 8)γDb; 9)γDc; 10)γEa; 11)γEb; 12)γEc; 13)γFa; 14)γFc. Using the same approach of the previous chapter for model selection we can now apply subset regression to the model 7.2.5 and compare the different models using the Schwartz information criterion. We apply subset regression just to constructed "current" variables $\boldsymbol{U}_t$ and "lagged" variables $\boldsymbol{V}_t$, leaving in each equation all the lagged canonical errors, $\boldsymbol{W}_t$.
Table 7.3 shows the likelihood and the values of the information criteria for the above

models, including the saturated model, for which the likelihood is identical to that of the canonical ARMA(1,1) fitted in chapter 4.

Table 7.2: Comparisons of the different models.

model	par.	lik	AIC	SIC	HIC
saturated	119	-3.00	-1563.3	-1040.1	-278.83
γAa	79	-2.54	-1363.4	-1016.1	-510.72
mod. fig. 7.3	73	-2.53	-1371.7	-1050.7	-583.73
γAc	79	-0.65	-233.4	113.97	619.33
γBa	79	-2.52	-1351.5	-1004.1	-498.78
γBc	79	-0.63	-221.44	-125.91	631.27
γCa	79	-2.54	-1364.7	-1017.4	-512.03
mod. fig. 7.2	72	-2.52	-1373.7	-1057.1	-596.51
γCc	79	-0.65	-234.69	112.66	618.02
γDa	79	-2.53	-1357.8	-1014.4	-505.07
γDb	79	-2.53	-1358.1	-1010.7	-505.34
γDc	79	-0.64	-227.73	119.62	624.98
γEa	79	-2.53	-1358.3	-1010.9	-505.57
γEb	79	-2.53	-1358.6	-1011.2	-505.85
γEc	79	-0.64	-228.23	119.12	624.48
γFa	79	-2.52	-1351.5	-1004.1	-498.78
γFc	79	-0.63	-221.45	125.91	631.27

For the same reasons already expressed in chapter 5 and chapter 6 we decided to follow the SIC. Eventually with this procedure we selected the model in figure 7.2. It includes

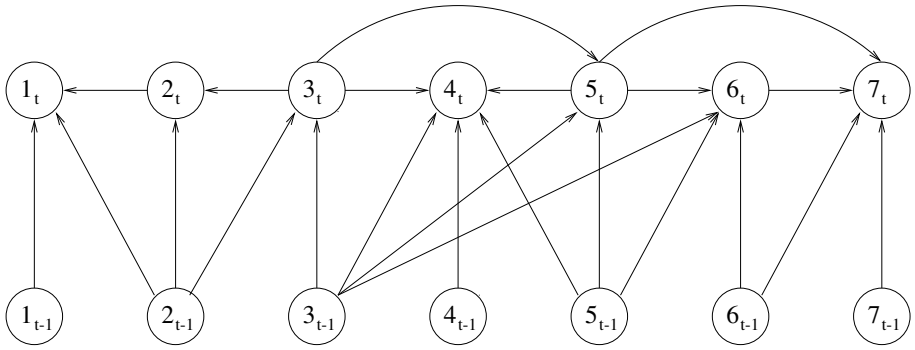


Figure 7.2: Model selected using the Schwartz information criterion.

the subgraph γCa, which resulted in a more adequate model than any other structure, for current variables.

Table 7.3 lists the coefficients in the model, the standard errors and t-values for the links shown in figure 7.2.

Table 7.3: Coefficients, standard errors and t-values of the links of DAG in figure 7.2.

Link	Coeff.	S.E.	t-value
$2_t \to 1_t$	0.8113	0.0240	33.8353
$1_{t-1} \to 1_t$	0.9935	0.0078	127.7258
$2_{t-1} \to 1_t$	-0.8133	0.0268	-30.3168
$3_t \to 2_t$	0.0954	0.0095	10.0242
$2_{t-1} \to 2_t$	0.9187	0.0084	108.9734
$2_{t-1} \to 3_t$	-0.0402	0.0095	-4.2368
$3_{t-1} \to 3_t$	1.0411	0.0108	96.4395
$3_t \to 4_t$	0.5517	0.0221	24.9848
$5_t \to 4_t$	0.4265	0.0243	17.5814
$3_{t-1} \to 4_t$	-0.4304	0.0285	-15.1016
$4_{t-1} \to 4_t$	0.7852	0.0348	22.5798
$5_{t-1} \to 4_t$	-0.3267	0.0288	-11.3299
$3_t \to 5_t$	0.8049	0.0175	45.9834
$3_{t-1} \to 5_t$	-0.8003	0.0177	-45.2444
$5_{t-1} \to 5_t$	0.9951	0.0072	137.9246
$5_t \to 6_t$	0.9360	0.0103	91.1074
$3_{t-1} \to 6_t$	-0.0150	0.0051	-2.9482
$5_{t-1} \to 6_t$	-0.8759	0.0208	-42.0782
$6_{t-1} \to 6_t$	0.9656	0.0136	70.9198
$5_t \to 7_t$	-0.0407	0.0077	-5.3111
$6_t \to 7_t$	0.9590	0.0121	79.4118
$6_{t-1} \to 7_t$	-0.8908	0.0144	-61.8128
$7_{t-1} \to 7_t$	0.9680	0.0089	108.6033

Table 7.4 shows the residual variances for the equations explaining the variables $1, \ldots, 7$. By fitting the equation 7.2.5 we have estimated efficiently the selected non-zero elements

Table 7.4: Residual variances.

Equation	1	2	3	4	5	6	7
Variance	0.0022	0.0055	0.0061	0.0004	0.0011	0.0003	0.0003

of $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Phi}_1$ in the hybrid model 7.1.4. The coefficient $\tilde{\boldsymbol{\Theta}}$ in 7.1.4 and 7.2.5 is a full matrix. Many of its (efficiently) estimated coefficients have small t-values. We proceed to estimate the coefficient $\boldsymbol{\Theta}_1$ in the fully structural VARMA(1,1) model 7.1.2. To do this we transformed $\boldsymbol{W}_t$ in 7.2.5 to $\boldsymbol{\Phi}_0\boldsymbol{W}_t$ using the matrix $\boldsymbol{\Phi}_0$ obtained from estimating the hybrid model. Because this results in a one to one transformation of $\boldsymbol{W}_t$, the likelihood of the model is unchanged, and so are the estimates of $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Phi}_1$. The estimate of the coefficient $\boldsymbol{\Theta}_1$ of the fully structural model appeared slightly simpler than that of $\tilde{\boldsymbol{\Theta}}$ in the hybrid model.

Our conclusion however is that the interpretation of the influence in the lagged residuals, whether canonical or structural, is less clear than the relationships revealed by $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Phi}_1$ and expressed in figure 7.2.

Observation of this graph indicates that variables 3, i.e. the two year interest rate, is the pivotal variable influencing contemporaneous interest rates. Its influence spreads down, through variables 2 and 1, and up, through variables 5, 6 and 7. Variable 4 is influenced by both 3 and 5, and 5 also directly influences 7.

The total number of parameters in the canonical VARMA(1,1) fitted in chapter 3, were 49 for the AR coefficients $\boldsymbol{\Phi}$, 49 for the MA coefficients and 21 elements of the covariance matrix of canonical residuals. The structural VARMA replaces the 49 AR coefficients and 21 covariances by 23 AR coefficients and 7 variances, a net reduction of 40 coefficients. This reduction supports the claim that the structural model gives a useful causal interpretation of the dependence between the variables.

A reasonable alternative model is shown in figure 7.3. It has been obtained by subregression from model γAa and shows a consistent flow of information from variable 3, which is the pivot variable, to the others. Its validity is supported by the values of the likelihood based methods of comparisons computed in table 7.3.

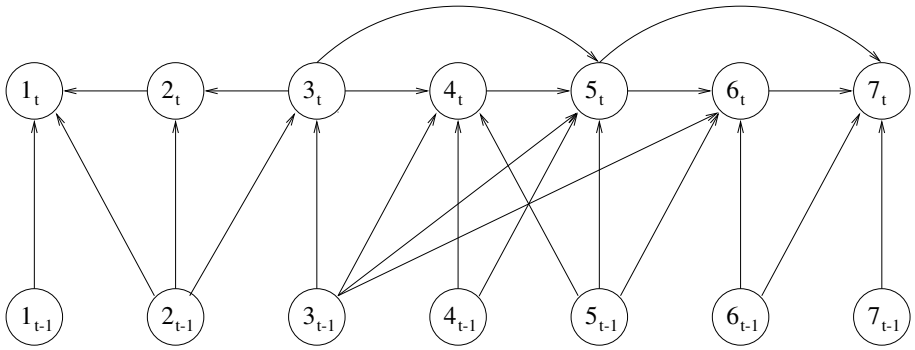


Figure 7.3: Alternative model.

7.4 Diagnostic checking

The basic assumption of the structural VARMA model is that the residual series are mutually incorrelated and are also white noise. To check this we evaluate the sample correlation matrix of the contemporaneous values and the sample lagged sample auto-correlations, or equivalent partial autocorrelations. A portmanteau test is to refer the scaled sum of squares of all the lagged correlations up to a chosen lag, say 20, to the chi-square distribution.

For the model in figure 7.2:

$$P = 600 \sum_{i=1}^{7} \sum_{j=1}^{7} \sum_{l=1}^{20} \tau_{ijl}^2 \;\sim\; \chi^2_{7\times 7\times 20-72}. \tag{7.4.1}$$

This is equivalent to the multivariate generalisation of the Box – Pierce test given by Hosking (1980). Such a test shows evidence of residual correlation. Also the observation of the contemporaneous correlation matrix, $\boldsymbol{R}$, in table 7.5. reveals the presence of

Table 7.5: Residual correlation matrix of model 7.2.

1	1.00						
2	-0.02	1.00					
3	0.14	0.47	1.00				
4	0.03	0.04	-0.01	1.00			
5	0.04	-0.02	-0.01	-0.00	1.00		
6	-0.02	-0.06	-0.06	0.07	0.11	1.00	
7	0.03	0.06	0.02	-0.01	0.05	-0.16	1.00
	1	2	3	4	5	6	7

significant correlation. The significant term $r_{1,3}$ is anticipated from the results of chapter 4 where we modelled the canonical residuals. To eliminate this and the other high values

we tentatively introduce new links into this model:

$$\begin{array}{llll} from & 3_t & to & 1_t \\ >> & 3_t & to & 7_t \\ >> & 3_{t-1} & to & 1_t \\ >> & 3_{t-1} & to & 2_t; \end{array}$$

emphasising in this way the role of the two years interest rate (variable 3). The corresponding DAG is represented in figure 7.4.

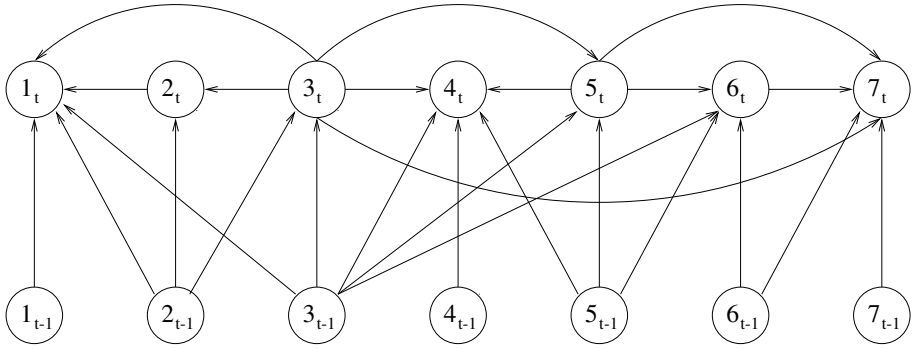


Figure 7.4: Model with added links.

Test 7.4.1 applied to residuals from such a model shows no remaining evidence of autocorrelation. The resulting residual correlation matrix is presented in table 7.6 which

Table 7.6: Residual correlation matrix of model 7.4.

1	1.00						
2	0.00	1.00					
3	0.00	-0.00	1.00				
4	0.04	0.06	-0.01	1.00			
5	0.04	-0.02	-0.01	-0.00	1.00		
6	-0.01	-0.03	-0.06	0.07	0.11	1.00	
7	0.04	0.05	0.03	0.00	0.01	-0.11	1.00
	1	2	3	4	5	6	7

shows just two values, $r_{5,6}$ and $r_{6,7}$, out of 21 exceeding the 2 standard error limits, which might be expected in any case. The new value of SIC (-1229.5) confirms that this is an improved model.

Chapter 8

Conclusions

This thesis has investigated aspects of the applications of graphical modelling to time series analysis.

We have considered the differences in modelling approach and inference which arise in this context when variables to which graphical modelling is applied are the current and lagged values of a single realisation of a univariate or multivariate time series. This includes in the case of ARMA models lagged values of the series innovations.

The main achievements are in the context of multiple time series models, particularly in the construction of structural VAR and VARMA models which we illustrate by two challenging real applications. In these we identify plausible mechanisms for their data generating process.

Structural models incorporate dependencies between contemporaneous variables. We have shown how graphical modelling can reduce sensibly the very large number of possible structural models which require consideration in the time series context.

In both our practical examples of the lending channel mechanism and the term structure the multivariate time series were highly correlated both contemporaneously and over time. The power of the graphical modelling approach to reveal dependencies in the context of highly correlated data has been successfully extended to the time series context.

References

Abadir, K.M., K. Hadri and E. Tzavalis (1999) The influence of VAR dimensions on estimator biases. *Econometrica*, **67**, 163–182.

Anderson, T.W. (1971) *The Statistical Analysis of Time Series.* Wiley: New York.

Ansley, C.F. and R. Kohn (1983) Exact likelihood of vector autoregressive process with missing or aggregated data. *Biometrika*, **70**, 275–278.

Ansley, C.F. and P. Newbold (1979) Multivariate partial autocorrelations. *ASA Proceeding of the Business and Economic Statistics Section*, 349–353.

Azzalini, A. (1996) *Statistical Inference Based on the Likelihood.* Chapman and Hall: London.

Bagliano, F.C. and A.C. Favero (forthcoming) "Il canale del credito della politica monetaria. Il caso Italia." in G. Vaciago (ed.) *Moneta e Finanza.* Il Mulino: Bologna.

Bernanke, B.S. and A.S. Blinder (1988) Credit, money and aggregate demand. *American Economic Review: Papers and Proceedings*, **78**, 435–439.

Box, G.E.P. and G.M. Jenkins (1976) *Time Series Analysis, Forecasting and Control*, revised edition. Holden-Day: Oakland.

Box, G.E.P. and G.C. Tiao (1977) A canonical analysis of multiple time series. *Biometrika*, **64**, 355–365.

Brockwell, P.J. and R.A. Davis (1991) *Time Series: Theory and Methods*, second edition. Springer-Verlag: New York.

Buttiglione, L. and G. Ferri (1994) Monetary policy transmission via lending rates in Italy: any lessons from recent experience? *Temi di Discussione*, **224**, Banca d'Italia.

Chatfield, C. (1996) *The Analysis of Time Series: an Introduction*, fifth edition. Chapman and Hall: London.

Cox, D.R. and H.D. Miller (1965) *The Theory of Stochastic Processes.* Chapman and Hall: London.

Diestel, R. (1997) *Graph Theory.* Springer-Verlag: New York.

Diggle, P.J. (1990) *Time Series: A Biostatistical Introduction.* Oxford University Press: Oxford.

Doob, J.L. (1953) *Stochastic Processes.* Wiley: New York.

Durbin, J. (1960) The fitting of time series models. *Review of the International Statistical Institute*, **28**, 233–244.

Greene, W.H. (1993) *Econometric Analysis*, second edition. Prentice-Hall: Englewood Cliffs.

Grimmett, G.R. and D.R. Stirzaker (1992) *Probability and Random Processes*, second edition. Oxford University Press: Oxford.

Hamilton, J.D. (1994) *Time Series Analysis.* Princeton University Press: Princeton.

Hannan, E.J. and B.G. Quinn (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society Series B*, **41**, 190–195.

Hannan, E.J. (1970) *Multiple Time Series.* Wiley: New York.

Harvey, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press: Cambridge.

Harvey, A.C. (1981) *Time Series Models.* Philip Allan: Oxford.

Hendry, D.F. (1995) *Dynamic Econometrics.* Oxford University Press: Oxford.

Hillmer, S.C. and G.C. Tiao (1979) Likelihood function of stationary multiple autoregressive moving average models. *Journal of American Statistical Association*, **74**, 652–660.

Hosking, J.R.M. (1980) The multivariate portmanteau statistics. *Journal of American Statistical Association*, **75**, 602–608.

Huang, D.S. (1970) *Regression and Econometric Methods.* Wiley: New York.

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.C. Lee (1985) *The Theory and Practice of Econometrics*, second edition. Wiley: New York.

KASHYAP, A.K. AND J.C. STEIN (1993) Monetary policy and bank lending. *NBER Working Paper*, **4317**.

LAURITZEN, S.L. AND D.J. SPIEGELHALTER (1988) Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society Series B*, **50**, 157–224.

LJUNG, G. AND G.E.P. BOX (1979) The likelihood function of stationary autoregressive-moving average models. *Biometrika*, **66**, 265–270.

LUCEÑO, A. (1994) A fast algorithm for the exact likelihood of stationary and partially nonstationary vector autoregressive moving average processes. *Biometrika*, **81**, 555–565.

LÜTKEPOHL, H. (1993) *Introduction to Multiple Time Series Analysis.* Springer-Verlag: Berlin.

MALLOWS, C.L. (1967) Linear processes are nearly gaussian. *Journal of Applied Probability*, **4**, 313–329.

MARDIA, K.V., J.T. KENT AND J.M. BIBBY (1979) *Multivariate Analysis.* Academic Press: London.

MARQUARDT, D.W. (1963) An algorithm for least squres estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, **11**, 431–441.

MAURICIO, J.A. (1995) Exact maximum likelihood estimation of stationary vector ARMA models. *Journal of American Statistical Association*, **90**, 282–291.

MIRON, J.A., C.D. ROMER AND D.N. WEIL (1993) Historical perspectives on the monetary transmission mechanism. *NBER Working Paper*, **4326**.

MONTI, A.C. (1998) A proposal for estimation of the parameters of multivariate moving-average models. *Journal of Time Series Analysis*, **19**, 209–219.

NEWTON, H.J. (1988) *Timeslab: a Time Series Analysis Laboratory.* Wadsworth and Brooks/Cole: Pacific Groove.

NICHOLLS, D.F. AND A.D. HALL (1979) The exact likelihood function of multivariate autoregressive moving average models. *Biometrika*, **66**, 259–264.

OSBORN, D.R. (1977) Exact and approximate maximum likelihood estimators for vector moving average process. *Journal of the Royal Statistical Society Series B*, **39**, 114–118.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman Publishers: San Mateo.

Phadke, M.S. and G. Kedem (1978) Computation of the exact likelihood function of multivariate moving average models. *Biometrika*, **65**, 511–519.

Piccolo, D. (1990) *Introduzione all'Analisi delle Serie Storiche*. NIS: Roma.

Piccolo, D. and G. Tunnicliffe Wilson (1984) A unified approach to ARMA model identification and preliminary estimation. *Journal of Time Series Analysis*, **5**, 183–204.

Priestley, M.B. (1981) *Spectral Analysis and Time Series.* Academic Press: London.

Reinsel, G.C. (1993) *Elements of Multivariate Time Series Analysis*. Springer-Verlag: New York.

Reinsel, G.C., S. Basu and S. Fwe Yap (1992) Maximum likelihood estimators in the multivariate autoregressive moving-average model from a generalized least squares viewpoint. *Journal of Time Series Analysis*, **13**, 133–145.

Rudin, W. (1976) *Principles of Mathematical Analysis*, third edition. McGraw-Hill: New York.

Schwartz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126.

Spirtes, P., C. Glymour and R. Scheines (1993) *Causation, Prediction and Search*. Springer-Verlag: New York.

Swanson, N.R. and C.W.J. Granger (1997) Impulse response functions based on a causal approach to residual orthogonalization in vector autoregression. *Journal of the American Statistical Association*, **92**, 357–367.

Tiao, G.C. and G.E.P. Box (1981) Modeling multiple time series with applications. *Journal of American Statistical Association*, **76**, 802–816.

Tiao, G.C. and R.S. Tsay (1989) Model specification in multivariate time series. *Journal of the Royal Statistical Society Series B*, **51**, 157–213.

Tunnicliffe Wilson, G. (1992) Structural models for structural change. *Quaderni di Statistica e Econometria*, **14**, 63–77.

Tunnicliffe Wilson, G. (1984) "Time series" in E.H. Lloyd (ed.) *Handbook of Applicable Mathematics.* Wiley: Chichester.

Tunnicliffe Wilson, G. (1973) The estimation of parameters in multivariate time series models. *Journal of the Royal Statistical Society Series B*, **35**, 76–85.

Whittaker, J.C. (1990) *Graphical Models in Applied Multivariate Statistics.* Wiley: Chichester.

Whittle, P. (1983) *Prediction and Regulation by Linear Least-Squares Methods*, second edition. Basil Blackwell: Oxford.

Whittle, P. (1963) On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, **50**, 129–134.

Zellner, A. and F. Palm (1974) Time series analysis and simultaneous equation econometric models. *Journal of Econometrics*, **2**, 17–54.