Master's Thesis presented to obtain the Degree of
Diplomvolkswirt

# Optimal Model Selection with Application to Volatility Models

submitted to: Prof. Dr. Wolfgang Härdle
Humboldt University Berlin
Faculty of Economics and Business Administration
Department of Statistics and Econometrics

by: Danilo Mercurio
Propststraße 7
10178 Berlin

Hiermit erkläre ich, daß ich die vorliegende Arbeit allein und unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Danilo Mercurio

Berlin, August 1999

# Contents

**Introduction** **1**

**1 The Regression Analysis of Time Series** **7**

1.1 The general model . . . . . 7

1.2 The general model selection criterion . . . . . 11

**2 Autoregressive processes** **15**

2.1 Some basic concepts of time series analysis . . . . . 15

2.2 The statistical analysis of AR processes . . . . . 19

2.2.1 The least square estimator . . . . . 19

2.3 A data-driven order selection algorithm . . . . . 21

2.3.1 Simulation results . . . . . 24

**3 State space models** **32**

3.1 State space models . . . . . 33

3.2 The Kalman Filter . . . . . 35

3.2.1 The extended Kalman filter . . . . . 36

3.2.2 Parameter identification . . . . . 38

3.3 Optimising the filter . . . . . 42

**4 Volatility models** **44**

4.1 Stylised facts of financial time series . . . . . 44

4.2 Volatility models . . . . . 50

4.2.1 Exponential-ARCH . . . . . 51
4.2.2 EGARCH . . . . . 52
4.2.3 Stochastic volatility (SV) . . . . . 53
4.3 Empirical evidence . . . . . 55
4.3.1 The problem of missing observations . . . . . 55
4.3.2 Model selection in practice . . . . . 57
4.3.3 A graphical example: the NIKKEI stock index . . . . . 64
4.3.4 Forecast confidence bands . . . . . 69

**Conclusion** **73**

**Bibliography** **75**

# List of Figures


1 NIKKEI stock index (upper plot) and the standardised returns (lower plot). . . . . 2

2 Zoom of the forecast confidence bands for the NIKKEI stock index . 4

2.1 Empirical density of $\hat{p}$ for the process (2.17) for different sample sizes, $M = 2p^{max}$ . . . . . 26

2.2 Empirical density of $\hat{p}$ for the process (2.18) for different sample sizes, $M = 2p^{max}$ . . . . . 27

2.3 The values of $p^{opt}$ for the process (2.19) computed by Monte Carlo according to formulae (2.20) the first, and (2.21) the last four . . . . 31

4.1 NIKKEI stock index (upper plot) and the standardised returns (lower plot). . . . . 45

4.2 Autocorrelations of the absolute values of the returns, of the squared and of the log-squared returns of the NIKKEI stock index. . . . . 48

4.3 Empirical densities of a standard normal sample (dotted) and of the standardised returns of the NIKKEI stock index (straight line). . . . 49

4.4 Forecasted log-volatility of the returns of the NIKKEI stock index . . 65

4.5 Estimations of the parameters of the stochastic volatility model for the NIKKEI stock index . . . . . 66

4.6 Estimations of the parameters of the EGARCH model for the NIKKEI stock index . . . . . 67

4.7 Autocorrelations of the log-squared returns minus the forecasted log-volatility for the NIKKEI stock index . . . . . 68
4.8 Zoom of the forecast confidence bands for the NIKKEI stock index . 72

# List of Tables


2.1 Risk Ratios . . . . . 29

4.1 Common initial conditions . . . . . 58
4.2 Specific initial conditions . . . . . 58
4.3 Model selection for the Japanese Yen/US-Dollar exchange rate . . . 61
4.4 Model selection for the ECU/US-Dollar exchange rate . . . . . 61
4.5 Model selection for the Standard and Poor index . . . . . 62
4.6 Model selection for the NIKKEI index . . . . . 62
4.7 Model selection for the Volkswagen stock prices . . . . . 63
4.8 Frequence of the observations which do not lie within the confidence bands relative to the whole sample . . . . . 71

# Introduction

This study is concerned with the question of model selection and forecasting for time series. First a general approach is presented, which can be applied in many different settings. Then attention is focused on autoregressive processes and on state space models, which build a very useful framework for the analysis of financial data. The problem of modelling the financial time series is finally addressed and some practical applications regarding the selection of the optimal forecasting model are shown.

Model selection is one of the most interesting topic in time series analysis. A series of observations of a certain quantity $\{y_t\}_{t=1}^{T}$ has been collected: for example the daily values of the NIKKEI stock index (see Figure 1). The aim of the researcher is to construct a model which explains the evolution of $y_t$ over time and possibly provides good forecasts of its future values. In general prediction is based on the past observations and as soon as a new observation becomes available the model is checked and eventually updated in order to optimise the forecasting performance.

The true data generating process is basically unknown and a very large number of models exist which could be taken into consideration for a particular problem. As a consequence in the first stage of model selection the practitioner, relying on his experience and knowledge of the statistical and economic theory usually limits his

attention to few particular models, that may suit his purposes.

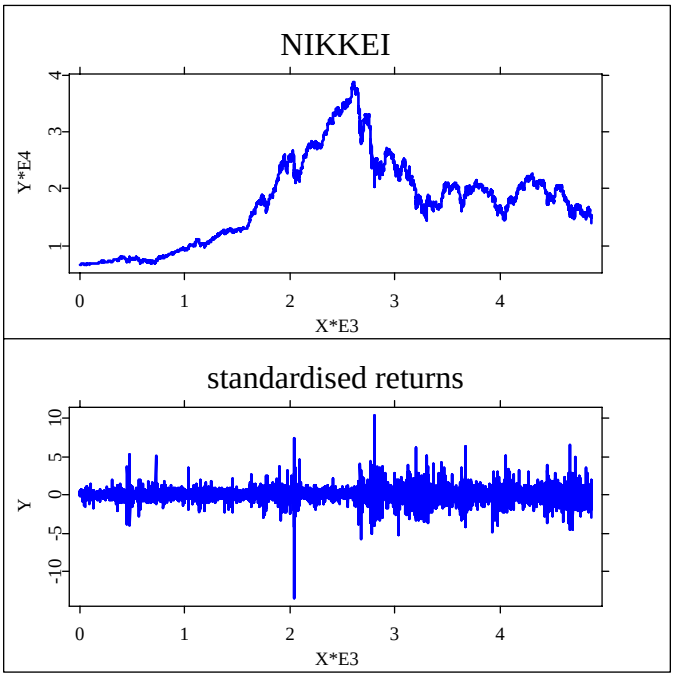


Figure 1: NIKKEI stock index (upper plot) and the standardised returns (lower plot).

In particular the models of the ARMA class are able to capture many features of economic time series (Lütkepohl (1993), Hamilton (1994), Hendry (1995) and Harvey (1992)) and are therfore very popular, although models which explicitly express the time series as a sum of unobserved components, the so called structural models (Harvey 1989) are also possible candidates. Nevertheless such models show very poor performances if applied to financial time series (Gouriéroux (1997), Hafner (1998), Bollerslev, Chou & Kroner (1992) and Engle (1995b)). Financial asset prices appear to follow a random walk process and are therefore basically unpredictable. Their growth rates, the returns are then modelled as a zero mean uncorrelated process, but

they exhibit a non-constant variance and a leptokurtic density if compared with the standard normal one. To cope with these features the volatility models have been proposed which express the conditional variance as a stochastic process itself.

Since the appearance of the first article on this topic by Engle (1995a) and the introduction of the ARCH processes, the literature on volatility models has become huge and many kind of processes have been developed to cope with the features of the volatility of the financial time series: GARCH (Bollerslev 1995), EGARCH (Nelson 1995) and stochastic volatility (SV) models (Harvey, Ruiz & Shephard 1995) probably represent the most famous and basic examples among them. Indeed, modelling and predicting the volatility is of essential importance for the financial praxis, for example in option prices (Hull & White (1987), Scott (1987), Johnson & Shanno (1987) and Härdle & Hafner (1997)), hedging (Feldmann (1998) and Gouriéroux (1997, Chapter 7)), portfolio analysis (Gouriéroux (1997, Chapter 9) and Hafner & Herwartz (1997)) and also for the appropriate construction of prediction intervals (see Figure 2 and Bollerslev (1995)).

As far as the models of the ARMA class are concerned, the model selection is usually done by the minimisation of one or several criteria that have been developed for this purpose: the Akaike, the Schwarz and the Hannan-Quinn criteria (see Lütkepohl (1993, Chapter 4) and Basci & Zaman (1998) for a recent survey). These criteria are constructed by taking into consideration a function of the estimated variance of the innovations $\hat{\sigma}$. The last two of them are consistent under quite general conditions of ergodicity and stationarity, while the Akaike criterion minimises the one step forecast error variance. If the aim of the researcher is to produce reliable forecasts, asymptotically consistence may not be so interesting and a criterion which minimises the

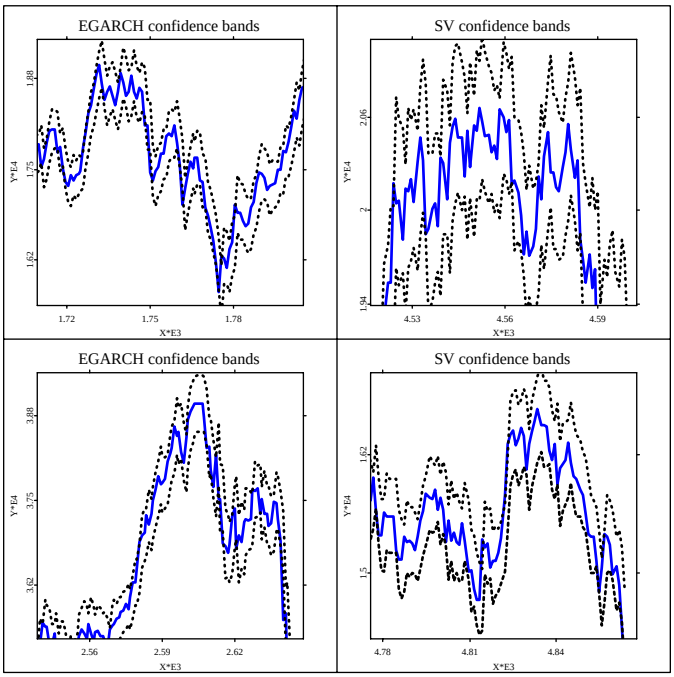


Figure 2: Zoom of the forecast confidence bands for the NIKKEI stock index

forecast uncertainty may be preferable, nevertheless the Akaike criterion shows the tendency of overestimating the true order in finite samples. This may lead to poor forecasting because of the variability of the resulting estimator.

For these reasons another criterion will be analysed which is able to work under general conditions and which can compare different modelling strategies. In practice at each time point the forecast for the next period, based only on the past data, is calculated, then the sum of the squared forecast error for each model is evaluated, and the model, which minimises this sum, is selected. The idea underlying the construction of this criterion is to adapt the cross validation principle to the context of time series: for independent data only one observation is usually dropped, but for a

time series one has to leave out all the subsequent observations.

The motivation of this model selection strategy is very intuitive and heuristically very easy to explain: the model which has shown in the past the best forecasting results is chosen, because it is expected to provide good predictions in the future, too. Unfortunately the theoretical properties of this criterion have not been derived yet, and we are able to show only some, but very promising results on simulated and real data. In any case the comparison of the out-of-sample forecast errors represents a common practice for the evaluation of the forecasting ability among different models; examples may be found in Franses & Dijk (1996) and in almost every issue of the *Journal of Forecasting*.

In the first chapter the general framework of the regression for time series and model selection is presented: the approach to this problem is derived from nonparametric statistical theory and it stresses the lack of knowledge about the true model. A model selection criterion, which focuses on the forecasting efficiency is defined for the general case.

In the second chapter autoregressive processes are analysed; these processes are very popular because they can approximate under general condition a wide class of stochastic processes (Lütkepohl 1986), and they are very easy to estimate. Some theoretical results about estimation forecast and model selection are shown and simulation results are presented which support our model selection strategy.

In the third chapter the wider class of state space models, which includes autoregressive processes as a special case, is considered; this kind of models is very useful

because it focuses on the concept of hidden process and therefore it applies to the analysis of the volatility process. We present an extended version of the Kalman filter algorithm (Chui & Chen (1998, Chapter 8), Singer (1998) and Elliot, Aggoun & Moore (1995, Chapter 6)) which allows the recursive estimation of the state process and of the parameters of a state space model, and therefore it provides an easy way of constructing the sum of the square forecast errors, and selecting among different ways of modelling the state process.

In the last chapter some practical application to real data is shown. Three modelling strategies: EARCH, EGARCH and SV are applied to daily financial time series. Consistently with the theory the EARCH model is usually outperformed. There is no regular pattern in the relative performance of EGARCH and SV, so that one has to distinguish from case to case.

## Acknowledgements

I am very obliged to thank Professor Wolfgang Härdle for his precious help and for giving me the possibility to accumulate many experiences in scientific research, and Professor Vladimir Spokoiny for his constant and very inspiring support.

Furthermore, I would like to thank some of the members of the Institute of Statistics and Econometrics for their helpful comments and their technical advises: Axel Werwatz, Marlene Müller, Rolf Tscherning, Christian Hafner, Ralph Brüggeman and Markus Krätzig.

Of course I am grateful to my parents for their patience and love.

# Chapter 1

# The Regression Analysis of Time Series

In this chapter the general regression model for time series is presented and the topics of model selection, estimation and model checking are addressed. We particularly focus on the question optimal model selection. As a solution of this problem we present a criterion, which minimises the sum of the squared forecasting residuals; this criterion is very general and it allows to choose the model with the best forecasting performance among models of very different nature.

## 1.1 The general model

Consider the following equation:

$$y_t = f_t + \varepsilon_t, \tag{1.1}$$

where the $\varepsilon_t$ are errors with $\varepsilon_t \sim (0, \sigma^2)$ and:

$$\mathrm{E}\left(y_t | y_{t-1}, y_{t-2}, \ldots, y_{t-n}, \ldots\right) = f_t. \tag{1.2}$$

Obviously $f_t$ is not independent from the past errors $\varepsilon_{t-1},\ \varepsilon_{t-2}, \ldots,$ but it is assumed to be independent from the present and future errors $\varepsilon_t,\ \varepsilon_{t+1}, \ldots.$

From equation (1.2) it can be derived that $f_t$ represents the optimal forecast for $y_t$. The expected forecast error is zero:

$$\mathrm{E}\left(y_t - f_t\right) = \mathrm{E}\left(f_t + \varepsilon_t - f_t\right) = 0,$$

and the forecast error variance is $\sigma^2$, which is the minimum attainable level of forecast uncertainty:

$$\mathrm{E}\left(y_t - f_t\right)^2 = \mathrm{E}\left(f_t + \varepsilon_t - f_t\right)^2 = \sigma^2.$$

If $f_t$ is estimated the forecast becomes suboptimal and the forecast uncertainty is in general larger. A further problem for the forecaster is that it is in general unknown which model suits the process $y_t$ best, indeed many models may be consistent with the general features of the process, but they may not all have the same forecasting ability.

Let $\hat{f}_t$ be an **estimated forecast**, i.e. an estimator of $f_t$ which is based only on the observations up to time $t-1$, then:

$$\mathrm{E}\left(y_t - \hat{f}_t\right)^2 = \sigma^2 + \mathrm{E}\left(f_t - \hat{f}_t\right)^2. \tag{1.3}$$

The second term of the right side of equation (1.3) is the mean square error (MSE), the additional forecast uncertainty due to the estimation. As usual the MSE can be decomposed into squared bias and variance:

$$\mathrm{E}\left(f_t - \hat{f}_t\right)^2 = \left(f_t - \mathrm{E}\,\hat{f}_t\right)^2 + \mathrm{E}\left(\hat{f}_t - \mathrm{E}\,\hat{f}_t\right)^2. \tag{1.4}$$

In general the variance of the estimation tends to disappear with the sample size, but this is usually not true for the bias.

Model selection and estimation play a key role in the minimisation of the forecasting uncertainty. Indeed these two issues are deeply connected, because from one side estimation implies a precise identification of the model, and from the other side the estimation techniques may affect the model selection.

Suppose that $f_t$ can be identified up to a set of parameters $\psi$, so that (1.1) can be rewritten equivalently in the following form:

$$y_t = f_t(\psi) + \varepsilon_t.$$

Then $f_t(\psi)$ is defined as the true (parametric) model which underlines the data generating process of $y_t$. In this case the estimation of $f_t$ is equivalent to the estimation of the parameters of $\psi$. In particular many sets of parameters may exists which fulfil the above equation, but for an efficient estimation one is interested in a set $\psi^*$, whose number of elements is finite and small, i.e. as small as possible and much smaller than the number of observations. Therefore in this study, following a widespread convention, we will refer to $\psi^*$ as the true model.

The true model is normally unknown and most of the literature on model selection actually focuses on the problem of the true model: the criteria of Schwarz and Hannan-Quinn are probably the most famous examples of such a strategy. A deep treatment of this topic can be found in almost any textbook on time series analysis; for this study Harvey (1992), Lütkepohl (1993) and Hamilton (1994) were taken into particular consideration.

Nevertheless another approach to the question of model selection is possible: one can target not the true model, but the best approximation of the real process. This

perspective is of particular interest when the true model is too large or too complicated to be estimated, for example, when it is defined by an infinite number of parameters. In particular we are interested in forecasting, and therefore we want to select the model that minimises the mean square forecast error.

Let $\{y_t\}_{t=1}^T$ be an observed time series, which fulfils the hypothesis underlying equations (1.1) and (1.2), let $\psi \in \Psi$ be a set of parameters which identifies a possible approximation of $f_t$, and let $\Psi$ be the set of models among which the selection takes place; furthermore define $Y_T = [y_T, y_{T-1}, \dots, y_1]^\top$ as the stacked vector of the observations up to time $T$. Then the optimal forecasting model minimises the MSE due to the estimation:

$$\psi_T^{opt} = \arg\inf_{\psi \in \Psi} \mathrm{E}\left(f_T - \hat{f}_T(\psi, Y_{T-1})\right)^2, \tag{1.5}$$

where $\hat{f}_T(\psi, Y_{T-1})$ represents the estimator of $f_T$ given $\psi$ and the observations up to time $T-1$.

Note that in the above definition the optimal model depends on the number of observations which are used for the estimation, therefore, for different sample sizes it is possible to have different optimal models: *"the larger the sample size, the more complicated the model can become"*. Nevertheless it is assumed that the optimal model remains constant over some subsamples, or equivalently, that it does not change as fast as $t$. Under this assumption the definition of the optimal model can be also stated in this form:

$$\psi_T^{opt} = \arg\inf_{\psi \in \Psi} \sum_{t=M}^{T} \mathrm{E}\left(f_t - \hat{f}_t(\psi, Y_{t-1})\right)^2. \tag{1.6}$$

The concept of an optimal model is not so common in time series analysis, and therefore it is worthwhile to discuss this idea it in more detail. A similar approach of

modelling time series can be found in Lewis & Reinsel (1982), Lütkepohl (1986) and Lütkepohl (1993), where the asymptotic properties are considered for the estimator of a stationary autoregressive process of infinite order. Such a process can be only approximated by a finite order autoregressive model. Under general conditions the estimator for the approximated model is consistent if the fitted order $\tilde{p}(T)$ tends to infinity with the sample size $T$, but at a much slower rate: for example

$$\tilde{p}(T) = T^{1/\delta} \qquad \forall\, \delta > 3. \tag{1.7}$$

Here, we are not able to derive asymptotic properties, nevertheless to allow some statistical analysis of our procedure we have to demand a certain constancy of the optimal model $\psi_T^{opt}$. Under this perspective the requirement expressed by equation (1.6) is similar to the one in (1.7).

Simulation results for autoregressive processes sustain this assumption, but they represent a specific example of a very simple process and therefore they do not have any claim of generality. In the case of real data, and complex models this simplifying assumption may not be fulfilled over large subsamples.

## 1.2 The general model selection criterion

The optimal loss function of equation (1.6) cannot be implemented directly because it consists of an expectation and it contains the unknown quantity $f_t$, which is precisely what we want to estimate. For that reason we have to consider another expression, whose characteristics are very similar to the ones of the sum of the MSE's. The sum of the squared forecasting errors represents a possible score function, whose dynamics with respect to $\psi$ are very similar to the one of (1.6). In that case the estimator of

the optimal model is defined as follows:

$$\hat{\psi}_T = \arg\inf_{\psi\in\Psi} \sum_{t=M}^{T} \left(y_t - \hat{f}_t(\psi, Y_{t-1})\right)^2 . \tag{1.8}$$

It is easy to see that the expectation of the sum of the squared forecasting errors coincides with the sum of the MSE's plus a constant.

Let us consider the score function and substitute (1.1) for $y_t$:

$$\sum_{t=M}^{T} \left(y_t - \hat{f}_t(\psi, Y_{t-1})\right)^2 = \sum_{t=M}^{T} \left(\varepsilon_t + \left(f_t - \hat{f}_t(\psi, Y_{t-1})\right)\right)^2 \tag{1.9}$$

The right side of equation (1.9) can be decomposed as follows:

$$\sum_{t=M}^{T} \left(\varepsilon_t + \left(f_t - \hat{f}_t(\psi, Y_{t-1})\right)\right)^2 =$$
$$\sum_{t=M}^{T} \varepsilon_t^2 \quad + \quad 2\sum_{t=M}^{T} \left(f_t - \hat{f}_t(\psi, Y_{t-1})\right) \varepsilon_t \quad + \quad \sum_{t=M}^{T} \left(f_t - \hat{f}_t(\psi, Y_{t-1})\right)^2 \tag{1.10}$$

The first term on the right side of (1.10), $\sum_{t=M}^{T} \varepsilon_t^2$, does not influence the choice of the model because it does not depend on $\psi$. It is the same for any estimator.

Let's consider now the expectation of the other two terms. Since $\hat{f}_t(\psi, Y_{t-1})$ is **independent** from $\varepsilon_t$, the cross term has zero expected value:

$$\mathrm{E}\left[2\sum_{t=M}^{T} \left(f_t - \hat{f}_t(\psi, Y_{t-1})\right) \varepsilon_t\right] = 2\,\mathrm{E}\left[\sum_{t=M}^{T} \left(f_t - \hat{f}_t(\psi, Y_{t-1})\right) \mathrm{E}\,\varepsilon_t \,|\mathcal{F}_{t-1}\right] = 0,$$

for that reason one can be confident that it essentially does not influence the model

selection, provided that its variance is small in comparison with the last term. Indeed:

$$\mathrm{E}\left[2\sum_{t=M}^{T}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)\varepsilon_t\right]^2 =$$

$$= 4\,\mathrm{E}\left[\sum_{t=M}^{T-1}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)\varepsilon_t+\left(f_T-\hat{f}_T(\psi,Y_{T-1})\right)\varepsilon_t\right]^2. \quad (1.11)$$

Consider the expansion of expression on the right side of the above equation:

$$4\,\mathrm{E}\left[\sum_{t=M}^{T-1}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)\varepsilon_t\right]^2 \quad (1.12)$$

$$+8\,\mathrm{E}\left[\sum_{t=M}^{T-1}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)\varepsilon_t\left(f_T-\hat{f}_T(\psi,Y_{T-1})\right)\varepsilon_T\right] \quad (1.13)$$

$$+4\,\mathrm{E}\left[\left(f_T-\hat{f}_T(\psi,Y_{T-1})\right)\varepsilon_T\right]^2. \quad (1.14)$$

Because of the **independence** of $\varepsilon_T$ from $\hat{f}_t(\psi,Y_{t-1})$ for any $t\le T$ the term (1.13) is zero:

$$8\,\mathrm{E}\left[\sum_{t=M}^{T-1}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)\varepsilon_t\left(f_T-\hat{f}_T(\psi,Y_{T-1})\right)\varepsilon_T\right] =$$

$$= 8\,\mathrm{E}\left[\sum_{t=M}^{T-1}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)\varepsilon_t\left(f_T-\hat{f}_T(\psi,Y_{T-1})\right)\mathrm{E}\,\varepsilon_T|\mathcal{F}_{T-1}\right]=0.$$

For the expression (1.14) we obtain:

$$\begin{aligned}4\,\mathrm{E}\left[\left(f_T-\hat{f}_T(\psi,Y_{T-1})\right)\varepsilon_T\right]^2 &= 4\,\mathrm{E}\left[\left(f_T-\hat{f}_T(\psi,Y_{T-1})\right)\mathrm{E}\,\varepsilon_T|\mathcal{F}_{T-1}\right]^2\\ &= 4\sigma^2\,\mathrm{E}\left[\left(f_T-\hat{f}_T(\psi,Y_{T-1})\right)\right]^2.\end{aligned}$$

After recursively applying the decomposition (1.11) to the term (1.12), one obtains the following expression for the variance of the cross term:

$$\mathrm{E}\left[2\sum_{t=M}^{T}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)\varepsilon_t\right]^2 = 4\sigma^2\,\mathrm{E}\left[\sum_{t=M}^{T}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)\right]^2.$$

It can be seen that the variance of the cross term is proportional to the expectation of the last term of the right side of equation (1.10), and therefore by the Chebyshev inequality it holds for any positive $\lambda$:

$$\mathrm{P}\left(\left|2\sum_{t=M}^{T}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)\varepsilon_t\right|\geq\lambda\right)\leq\frac{4\sigma^2\,\mathrm{E}\sum_{t=M}^{T}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)^2}{\lambda^2}.$$

The magnitude of the cross term is with high probability much smaller than the expectation of the last term. For these reasons we can say that the third term of (1.10) is the one that essentially drives the dynamics of (1.8) with respect to $\psi$, furthermore we know from the law of the large numbers for martingale (Jacod & Shiryaev 1987) that a sum of random variables tends to coincide with its expectation:

$$\frac{\sum_{t=M}^{T}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)^2}{\mathrm{E}\sum_{t=M}^{T}\left(f_t-\hat{f}_t(\psi,Y_{t-1})\right)^2}\xrightarrow{\mathrm{P}}1 \tag{1.15}$$

which is precisely the optimal score function. We may therefore expect that $\hat{\psi}_T$ from equation (1.8) represents a good estimator for the optimal model $\psi_T^{opt}$ defined in equation (1.6).

Note that while analysing the sum of the squared forecast error we have used the fact that the estimated forecast is only based on past observations, and therefore is independent from present and future errors. Similar results cannot be obtained if the whole data set contributes to the estimation and the residual sum of square is then computed.

# Chapter 2

# Autoregressive processes

In the first section of this chapter we briefly introduce the stochastic processes and some related concepts. Our attention then focuses on autoregressive processes, which posses the important property of approximating a wide class of stochastic processes. In the second section we consider the estimation of the parameters of autoregressive processes of unknown order. Finally the model selection algorithm is applied to the choice of the optimal order of an autoregressive process and simulation results are shown, which we interpret as evidence in favour of our algorithm.

## 2.1 Some basic concepts of time series analysis

Here some introductory ideas of time series analysis are briefly presented. In particular we take into consideration these results, which underline the importance of autoregressive processes as a mean to approximate the wide class of stationary time series. A deeper treatment of these topics can be found in Lütkepohl (1993) and Hamilton (1994).

A **stochastic process** is a function of two variables, a random variable $y_t$ and a time variable $t \in T \subset \mathbb{R}$. The random variable $y_t$ is a measurable function defined on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ where $\Omega$ is the set of all possible events $\omega$, $\mathcal{F}$ is the $\sigma$-field and P is the probability measure. A stochastic process can be represented as follows:

$$\{y_t(\omega); \quad t \in T\};$$

but for saving notation we will mostly use throughout the following study the standard notation: $y_t$.

If we fix an elementary event $\bar{\omega} \in \Omega$ the function $y_t(\bar{\omega})$ becomes a deterministic function of time and it is called a realization, a story of the process. On the other hand if we fix a particular $\bar{t} \in T$, then we obtain a random variable $y_{\bar{t}}(\omega)$ which is called a section of the process at time $\bar{t}$.

A stochastic process can be univariate or multivariate, in discrete or continuous time. It is called univariate if the generic section of the process is a random variable, multivariate if the section of the process is a random vector. It is in discrete time if $T$ is a countable set such as $\mathbb{Z}$ or $\mathbb{N}$. It is in continuous time if $T$ is an uncountable set such as $\mathbb{R}$. In this study we focus our attention only on discrete time processes.

Let $y_t$ be a stochastic process, $y_t$ is called a **weakly stationary process** if:

$$\mathrm{E}\, y_t = \mu \qquad \forall t \in T, \tag{2.1}$$

and

$$\mathrm{E}(y_t - \mu)(y_{t-h} - \mu) = \gamma_y(h) = \gamma_y(-h) \qquad \forall t \in T, h = 0, 1, 2, \ldots \tag{2.2}$$

Condition (2.1) means that all $y_t$ have the same finite mean $\mu$ and (2.2) requires that the autocovariances $\gamma_y(h)$ of the process do not depend on $t$ but just on the time period $h$ the two variables $y_t$ and $y_{t-h}$ are apart. Few economic time series are stationary, but empirical applications show that they can be made stationary, at least approximatively, by differencing.

The simplest form of stochastic process is probably represented by the **white noise process**. Let $\varepsilon_t$ be a stochastic process: $\varepsilon_t$ is called white noise if $\mathrm{E}\,\varepsilon_t = 0$, $\mathrm{E}\,\varepsilon_t^2 = \sigma_t^2$, $\mathrm{E}\,\varepsilon_t\varepsilon_s = 0 \quad \forall t \neq s$. A sequence of zero mean i.i.d. random variables is a recurrent example of a white noise process, but independence and identical distribution are not necessary conditions. Indeed, financial returns (see Chapter 4) are often modelled as uncorrelated but not independent, heteroskedastic white noise sequences.

The white noise process can be used to construct processes with more interesting dynamics, such as the **moving average process**. Let $y_t$ be a stochastic process and let $\varepsilon_t$ be a white noise process. Then $y_t$ is called a moving average process of order $q$, MA($q$) if it can be written as follows:

$$y_t = \mu + \varepsilon_t + \alpha_1\varepsilon_{t-1} + \ldots + \alpha_q\varepsilon_{t-q}. \tag{2.3}$$

It is very easy to show that an MA process is stationary.

Moving average processes are often very useful in the analysis of other stationary processes because of **Wold's Decomposition Theorem**. According to this theorem any stationary process $y_t$, can be written as the sum of two uncorrelated processes $d_t$ and $z_t$:

$$y_t = d_t + z_t,$$

where $d_t$ is a deterministic process and $z_t$ has an MA($\infty$) representation:

$$z_t = \sum_{i=0}^{\infty} \alpha_i \varepsilon_{t-i}.$$

**Autoregressive processes**, which have also been mentioned in the previous chapter, represents a very popular way of modelling time series in empirical applications. Let $y_t$ be a stochastic process and let $\varepsilon_t$ be a white noise process. Then $y_t$ is called an autoregressive process of order $p$, AR($p$) if it can be written in the following form:

$$y_t = \theta_0 + \theta_1 y_{t-1} + \ldots + \theta_p y_{t-p} + \varepsilon_t. \tag{2.4}$$

The persistence of the dynamics of an AR($p$), is determined by the roots of its characteristic polynomial. In particular, $y_t$ is **stable** if the following condition holds:

$$\theta_p x^p + \ldots + \theta_2 x^2 + \theta_1 x + \theta_0 \neq 0 \qquad \forall\ |x| \leq 1.$$

A stable AR process is also stationary, and is therefore **invertible**. Hence, it possesses an MA($\infty$) representation.

The analysis of the roots of the characteristic polynomial is also interesting for the MA process. If the following condition holds:

$$\alpha_q x^q + \ldots + \alpha_2 x^2 + \alpha x + 1 \neq 0 \qquad \forall\ |x| \leq 1,$$

the MA process is invertible and it possess an AR($\infty$) representation.

We can therefore conclude that almost any time series which is stationary, or can be made stationary by differencing can be approximated by a finite autoregressive process.

## 2.2 The statistical analysis of AR processes

The various propositions and conditions listed in the previous section show how interesting AR processes are from a statistical point of view: they can lead to parsimonious approximations of stationary time series. Furthermore, they are very easy to estimate.

### 2.2.1 The least square estimator

Here, the least squares (LS) estimator of an autoregressive process will be analysed. Consider equation (2.4) in a slightly different form:

$$y_t = \mathbf{x}_{t-1,p^*}^\top \boldsymbol{\theta} + \varepsilon_t, \tag{2.5}$$

with $\mathbf{x}_{t-1,p^*}^\top = [1 \; y_{t-1} \ldots y_{t-p^*}]$ and $\boldsymbol{\theta}^T = [\theta_0 \; \theta_1 \ldots \theta_{p^*}]$; suppose that $p^*$ is the true (minimal) order and that we observe a realisation $\{y\}_{t=1}^T$ of the process (2.4). Then the LS estimator for $\boldsymbol{\theta}$ is given by:

$$\hat{\boldsymbol{\theta}}_{p^*} = \left( \sum_{t=p^*+1}^{T} \mathbf{x}_{t-1,p^*} \mathbf{x}_{t-1,p^*}^\top \right)^{-1} \left( \sum_{t=p^*+1}^{T} \mathbf{x}_{t-1,p^*} y_t \right), \tag{2.6}$$

and the error of the estimation is given by:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{p^*} - \boldsymbol{\theta} &= \left( \sum_{t=p^*+1}^{T} \mathbf{x}_{t-1,p^*} \mathbf{x}_{t-1,p^*}^\top \right)^{-1} \left( \sum_{t=p^*+1}^{T} \mathbf{x}_{t-1,p^*} y_t \right) - \boldsymbol{\theta} \\ &= \left( \sum_{t=p^*+1}^{T} \mathbf{x}_{t-1,p^*} \mathbf{x}_{t-1,p^*}^\top \right)^{-1} \left( \sum_{t=p^*+1}^{T} \mathbf{x}_{t-1,p^*} \left( \mathbf{x}_{t-1,p^*}^\top \boldsymbol{\theta} + \varepsilon_t \right) \right) - \boldsymbol{\theta} \\ &= \left( \sum_{t=p^*+1}^{T} \mathbf{x}_{t-1,p^*} \mathbf{x}_{t-1,p^*}^\top \right)^{-1} \left( \sum_{t=p^*+1}^{T} \mathbf{x}_{t-1,p^*} \varepsilon_t \right). \end{aligned} \tag{2.7}$$

An analogous result holds if we estimate any $\hat{\boldsymbol{\theta}}_{p'}$ such that $p' > p^*$:

$$\hat{\boldsymbol{\theta}}_{p'} - \boldsymbol{\theta} = \left( \sum_{t=p'+1}^{T} \mathbf{x}_{t-1,p'} \mathbf{x}_{t-1,p'}^\top \right)^{-1} \left( \sum_{t=p'+1}^{T} \mathbf{x}_{t-1,p'} \varepsilon_t \right). \tag{2.8}$$

Note that whenever necessary we construct a conformable vector, whose auxiliary elements are neutral in an algebraic sense. In the previous case for example, the dimension of $\boldsymbol{\theta}$ is $p'+1$ and the last $(p'+1)-(p*+1)$ coefficients are set equal to zero.

In the standard regression analysis it is quite easy to derive the finite sample properties of the LS estimator, but in the context of time series analysis only asymptotic properties can be derived. In particular the LS estimator is only asymptotically unbiased.

Let $\text{plim}\left(T^{-1}\left(\sum_{t=p+1}^{T}\mathbf{x}_{t-1,p}\mathbf{x}_{t-1,p}^{\top}\right)\right)=V$ be finite and nonsingular. Then under general conditions (Hamilton 1994, Chapter 8), for $p'\geq p^*$, it holds that:

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{p'}-\boldsymbol{\theta})\xrightarrow{d}N(0,\sigma^2V^{-1}). \tag{2.9}$$

The above expression means that the LS estimator is asymptotically unbiased. One can also see that the dimension of the covariance matrix is $(p'+1)\times(p'+1)$ so that in general the estimation variance grows with the number of parameter to estimate.

The estimation error is on the contrary quite different if we estimate any $\hat{\boldsymbol{\theta}}_{p''}$, where $p''$ is smaller than $p^*$: implicitly all the coefficients $\theta_i$, for $p''<i\leq p^*$ are restricted to zero. Define $\boldsymbol{\theta}_{p''}$ and $\boldsymbol{\theta}_{-p''}$ as the partitions of the vector of the true parameters $\boldsymbol{\theta}$, which consist of the first $p''+1$ and the last $(p^*+1)-(p''+1)$ elements of $\boldsymbol{\theta}$ respectively. Then the estimation error of $\boldsymbol{\theta}_{-p''}$ is trivially:

$$\hat{\boldsymbol{\theta}}_{-p''}-\boldsymbol{\theta}_{-p''}=-\boldsymbol{\theta}_{-p''}. \tag{2.10}$$

The last coefficients are simply not estimated. Define the partitions of $\mathbf{x}_{t,p^*}$: $\mathbf{x}_{t,p''}$ and

$\mathbf{x}_{t,-p''}$, like the partitions of $\boldsymbol{\theta}$. Then the estimation error of $\boldsymbol{\theta}_{p''}$ is:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{p''} - \boldsymbol{\theta}_{p''} &= \left(\sum_{t=p''+1}^{T} \mathbf{x}_{t-1,p''}\mathbf{x}_{t-1,p''}^{\top}\right)^{-1} \left(\sum_{t=p''+1}^{T} \mathbf{x}_{t-1,p''} y_t\right) - \boldsymbol{\theta}_{p''} \\
&= \left(\sum_{t=p''+1}^{T} \mathbf{x}_{t-1,p''}\mathbf{x}_{t-1,p''}^{\top}\right)^{-1} \left(\sum_{t=p''+1}^{T} \mathbf{x}_{t-1,p''} \left(\mathbf{x}_{t-1,p^*}^{\top}\boldsymbol{\theta} + \varepsilon_t\right)\right) - \boldsymbol{\theta}_{p''} \\
&= \left(\sum_{t=p''+1}^{T} \mathbf{x}_{t-1,p''}\mathbf{x}_{t-1,p''}^{\top}\right)^{-1} \\
&\quad \left(\sum_{t=p''+1}^{T} \mathbf{x}_{t-1,p''} \left(\mathbf{x}_{t-1,p''}^{\top}\boldsymbol{\theta}_{p''} + \mathbf{x}_{t-1,-p''}^{\top}\boldsymbol{\theta}_{-p''} + \varepsilon_t\right)\right) - \boldsymbol{\theta}_{p''} \\
&= \left(\sum_{t=p''+1}^{T} \mathbf{x}_{t-1,p''}\mathbf{x}_{t-1,p''}^{\top}\right)^{-1} \left(\sum_{t=p''+1}^{T} \mathbf{x}_{t-1,p''} \left(\mathbf{x}_{t-1,-p''}^{\top}\boldsymbol{\theta}_{-p''}\right)\right) + \\
&\quad \left(\sum_{t=p''+1}^{T} \mathbf{x}_{t-1,p''}\mathbf{x}_{t-1,p''}^{\top}\right)^{-1} \left(\sum_{t=p''+1}^{T} \mathbf{x}_{t-1,p''} \varepsilon_t\right). \qquad (2.11)
\end{aligned}$$

As far as the asymptotic properties of the LS estimator for $p'' < p^*$ are concerned it is straightforward to see that $\hat{\boldsymbol{\theta}}_{p''}$ is not unbiased. The error expressed in (2.10) is not a random variable and it never goes to zero. Yet it is possibly very small, therefore it may happen that in finite samples the MSE due to $p''$ is smaller than the MSE due to $p^*$.

## 2.3 A data-driven order selection algorithm

The previous section underlines how the choice of the order may influence the results of the estimation. Here a Monte Carlo simulation concerning the choice of the optimal order for AR processes is performed. The theory that we developed in Chapter 1 can be implemented very easily.

If the parameters are known than the optimal one step forecast for $y_{t+1}$ is

$$\tilde{y}_{t+1|t} = \mathbf{x}_{t,p}^{\top}\boldsymbol{\theta}_p,$$

on the other hand if $\boldsymbol{\theta}_p$ is unknown we have a suboptimal forecast with estimated parameters:

$$\hat{y}_{t+1|t} = \mathbf{x}_{t,p}^{\top}\hat{\boldsymbol{\theta}}_{t,p},$$

where the expression $\hat{\boldsymbol{\theta}}_{t,p}$ defines the estimator of $\boldsymbol{\theta}_p$ based only on the observations up to time $t$.

The general model selection criterion defined in equation (1.5) minimises the forecast uncertainty due to the estimation. In the case of the AR processes it becomes:

$$p_T^{opt} = \arg\inf_{p\in\Pi} \mathrm{E}(\mathbf{x}_{t,p}^{\top}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_p))^2 \quad \text{with} \quad \Pi \subset \mathbb{N}. \tag{2.12}$$

The optimal order, $p_T^{opt}$ is not necessarily constant over time. For moderate samples a small order may be adequate, while for a large sample even a quite large order may be estimated with great precision.

Nevertheless we assume that $p_t^{opt}$ does not grow as fast as $t$, or equivalently that it is constant over some subsamples:

$$p_M^{opt} = p_{M+1}^{opt} = \ldots = p_{T-1}^{opt} = p_T^{opt} \qquad M < T. \tag{2.13}$$

We define a function $\Xi(p)$ analogous to the one in equation (1.6) such that:

$$p_T^{opt} = \arg\inf_{p\in\Pi} \Xi(p) = \arg\inf_{p\in\Pi} \sum_{t=M}^{T} \mathrm{E}(\mathbf{x}_{t,p}^{\top}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{t,p}))^2, \tag{2.14}$$

where $\hat{\boldsymbol{\theta}}_{t,p}$ is the running LS estimator of $\boldsymbol{\theta}$, which uses only the observation up to time $t$:

$$\hat{\boldsymbol{\theta}}_{t,p} = \left( \sum_{i=p+1}^{t} \mathbf{x}_{i-1,p} \mathbf{x}_{i-1,p}^{\top} \right)^{-1} \left( \sum_{i=p+1}^{t} \mathbf{x}_{i-1,p} y_i \right). \tag{2.15}$$

It must be noted that the computation of (2.15) does not require a matrix inversion at each step. Indeed one can update the estimator of $\boldsymbol{\theta}_p$ in a much simpler way:

$$\hat{\boldsymbol{\theta}}_{t,p} = \hat{\boldsymbol{\theta}}_{t-1,p} + (X_{t-1,p}^{\top} X_{t-1,p})^{-1} \mathbf{x}_t (y_{t+1} - \mathbf{x}_{t,p}^{\top} \hat{\boldsymbol{\theta}}_{t-1,p}) / e_t,$$

where:

$$X_{t,p}^{\top} = \begin{bmatrix} \mathbf{x}_{p+1,p} & \mathbf{x}_{p+2,p} & \dots & \mathbf{x}_{t-1,p} & \mathbf{x}_{t,p} \end{bmatrix},$$

and the inverse:

$$(X_{t,p}^{\top} X_{t,p})^{-1} = (X_{t-1,p}^{\top} X_{t-1,p})^{-1} - (X_{t-1,p}^{\top} X_{t-1,p})^{-1} \mathbf{x}_{t,p} \mathbf{x}_{t,p}^{\top} (X_{t-1,p}^{\top} X_{t-1,p})^{-1} / e_t,$$

and

$$e_t = 1 + \mathbf{x}_{t,p}^{\top} (X_{t-1,p}^{\top} X_{t-1,p})^{-1} \mathbf{x}_{t,p}.$$

The above formulae represent the recursive least square, which can be interpreted as a special case of the Kalman filter (see Chapter 3 and Harvey (1992, pagg. 98-100)).

As in the general case one cannot minimise $\Xi(p)$ because it contains the unknown quantity $\boldsymbol{\theta}$ and the expected value operator. Instead we must use a function which doesn't contain unknown quantities other than $p_T^{opt}$, and whose behaviour in $p$ is very similar to the one of $\Xi(p)$. For that purpose a natural choice is represented by:

$$S(p) = \sum_{t=M}^{T-1} \left( y_{t+1} - \mathbf{x}_{t,p}^{\top} \hat{\boldsymbol{\theta}}_t \right)^2,$$

which expresses the loss function of equation (1.8) for the case of AR processes. It follows that the estimator of $p_T^{opt}$ is:

$$\hat{p}_T = \arg\inf_{p \in \Pi} S(p). \tag{2.16}$$

For those quantities the general theory developed in Chapter 1 holds, because we are analysing a recursive estimator. Once more we want to underline the fact that the optimal order $p_t^{opt}$ may increase with the sample size, and obviously we are interested in its value at the end of the sample, $p_T^{opt}$, because the final aim is to produce good out of sample forecasts. It is clear that minimising $S(p)$ for very large $M$ increases the risk of underestimating $p_T^{opt}$. But for small values of $M$ one cannot appeal any more to the low of the large numbers and the approximation of the behaviour of $\Xi(p)$ through $S(p)$ stated in equation (1.15) is not very good.

### 2.3.1 Simulation results

The theoretical properties of the order selection criterion that we present have not been developed yet, so that we have to evaluate its performances with the help of a simulation study. Stationary AR(2) and AR(3) processes are taken into consideration. For each process we simulate 100 realisations, and for different sample sizes we analyse the performances of our order selection criterion. Particular attention is then given the choice of $M$.

Let us consider the following stationary autoregressive processes:

$$y_t = 0.7y_{t-1} - 0.12y_{t-2} + \varepsilon_t, \tag{2.17}$$

and

$$y_t = 1.2y_{t-1} - 0.47y_{t-2} + 0.06y_{t-3} + \varepsilon_t, \tag{2.18}$$

where $\varepsilon_t \sim N(0,1)$.

Figures (2.1) and (2.2) show the empirical distributions of $\hat{p}_T$ for (2.17) and (2.18) respectively, with $p^{min} = 1$ and $p^{max} = 5$. The values of $\hat{p}_T$ have been computed according to (2.16) and $M = 2p^{max}$ for any of the following sample sizes: $T = 100$, $T = 250$, $T = 500$ and $T = 1000$. The distributions show that the estimated optimal order $\hat{p}$ is in general smaller or equal to the true order $p^*$. Overfitting tends to be avoided, because the estimation of further parameters (which are in fact zero) cannot improve the forecasting performance, but it leads to a larger MSE. The estimator of the optimal order tends to be lower than $p^*$, in particular if some conditions are met, such as small sample size, and small absolute value of the coefficients of the most lagged values with respect to the ones of the most recent. The processes that we have generated show exactly this feature: $|\theta_1| > |\theta_2| > |\theta_3|$, but this characteristic is shared with many empirical series.

As expected, the optimal order tends to coincide with the true order for increasing sample size. We see indeed that $\hat{p}$ tends to $p^*$ if the number of observations is large enough, this seems to confirm that the true order $p^*$ represent an upper bound for the optimal order $p^{opt}$.

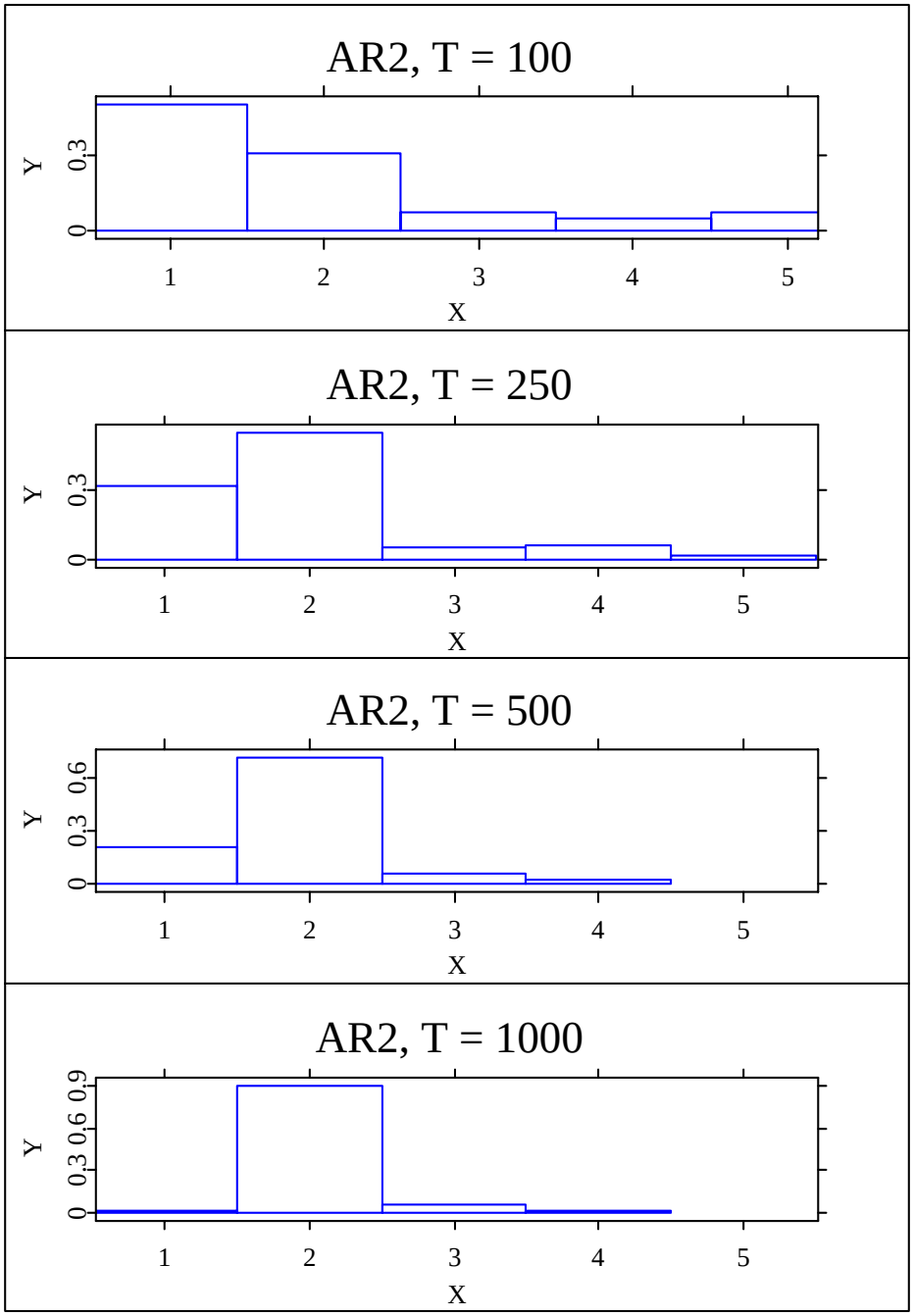


Figure 2.1: Empirical density of $\hat{p}$ for the process (2.17) for different sample sizes, $M = 2p^{max}$

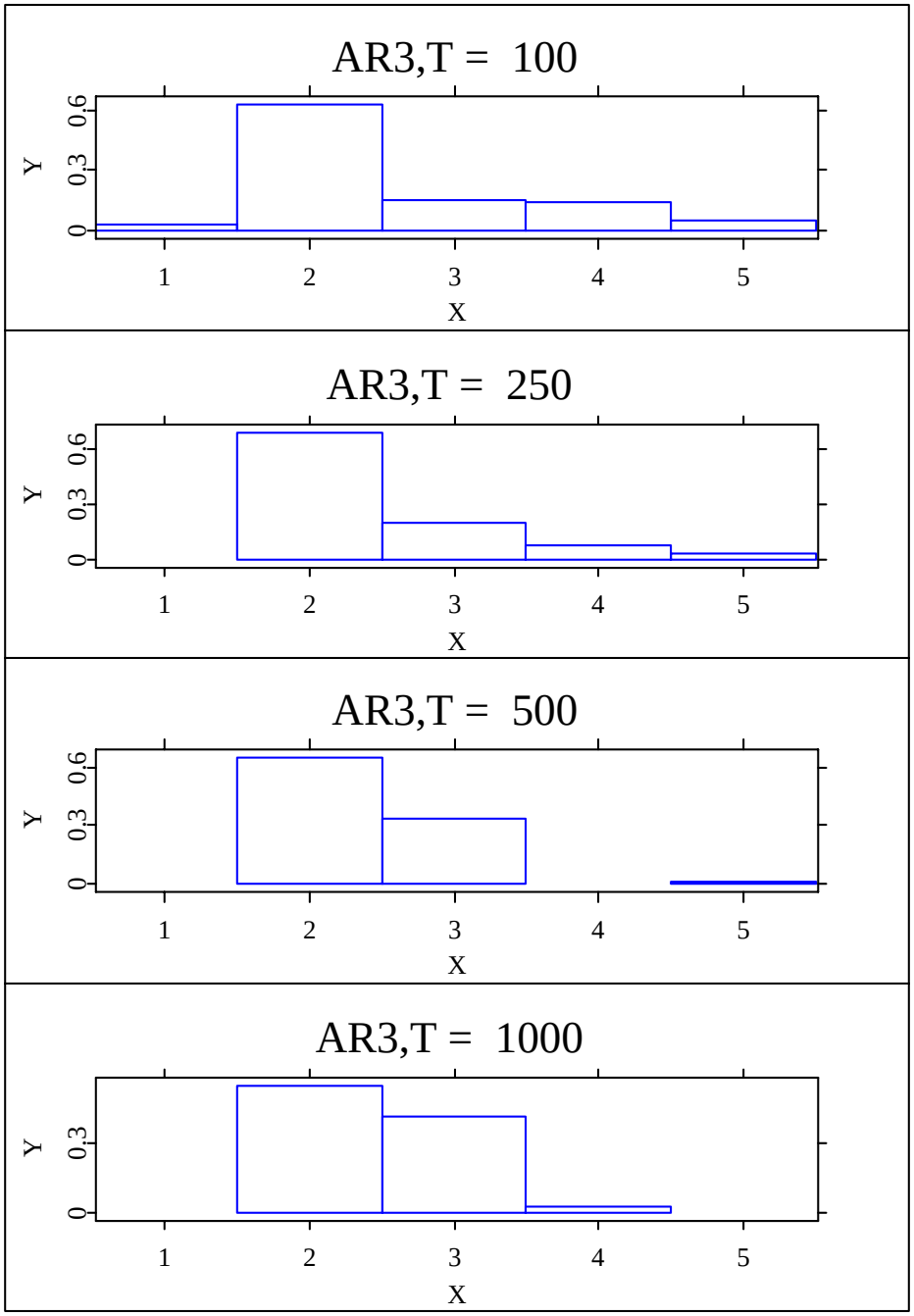


Figure 2.2: Empirical density of $\hat{p}$ for the process (2.18) for different sample sizes, $M = 2p^{max}$

To gain a better feeling of the efficiency of $\hat{p}_T$, the estimator of the optimal order $p_T^{opt}$, we have to compare the estimated order with the optimal order. The latter cannot be computed explicitely and therefore Monte Carlo simulations are needed. We perform 100 realisations, each time generating 500 observations of the following AR(3) process:

$$y_t = 1 + 1.6y_{t-1} - 0.74y_{t-2} + 0.105y_{t-3} + \varepsilon_t, \tag{2.19}$$

where $\varepsilon_t \sim N(0,1)$. The optimal order $p_{t,MC}^{opt}$ (Figure 2.3 upper plot) can be therefore calculated as the minimising argument of the mean among all realisations at a specific time $t$:

$$p_{t,MC}^{opt} = \arg\inf_{p\in\Pi} 100^{-1} \sum_{\omega=1}^{100} \left(\mathbf{x}_{p,t}^{\top}(\omega)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{p,t}\right)\right)^2, \tag{2.20}$$

here the index $\omega$ is related to the particular realisation of the process and the subscript MC stands for Monte Carlo. One can see that the optimal order is 2 for $t \lesssim 100$, and 3 for $t \gtrsim 100$. In a neighbourhood of $t = 100$ it is not clear if the optimal is order 2 or 3. Certainly 100 realisations are not enough to distinguish exactly at which point the optimal order changes from 2 to 3, but it is also possible that over a certain interval the values of the risk of order 2 is equal or at least very similar to the value of the risk of order 3, so that one is indifferent between the two orders.

Now we want to consider the implication of the choice of the interval $M$ for the estimation of $p_t^{opt}$. We consider the Monte Carlo version of formula (2.14):

$$p_{T,MC}^{opt} = \arg\inf_{p\in\Pi} \Xi_{MC}(p) = \arg\inf_{p\in\Pi} 100^{-1} \sum_{t=M}^{T} \sum_{\omega=1}^{100} \left(\mathbf{x}_{p,t}^{\top}(\omega)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{p,t}\right)\right)^2. \tag{2.21}$$

The four lower plots of figure (2.3) show the values of $p_{T,MC}^{opt}$ according to the above formula, for different values of $M$. The sample size starts with $T = 50$ and then

increases with steps of 50. It can be seen that all these plots give a good approximation of the upper plot, because the optimal order is indeed constant over some intervals. Nevertheless some problems may arise if one chooses a value $M$ too small in comparison to the sample size. In that case the the spring from $\hat{p} = 2$, to $\hat{p} = 3$ may happen with some delay and the optimal order may be underestimated.

To analyse the efficiency of the order selection criterion the following index of risk is constructed:

$$\frac{100^{-1} \sum_{t=M}^{T} \sum_{\omega=1}^{100} \left( \mathbf{x}^{\top}_{\widehat{p(\omega)},t}(\omega) \left( \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\widehat{p(\omega)},t} \right) \right)^2}{\Xi_{MC}(p_T^{opt})}, \tag{2.22}$$

where the numerator represents the average of the square errors of each realisation, and $\widehat{p(\omega)}$ is the estimated order for that realisation $y_t(\omega)$ according to equation (2.16). In practice the index (2.22) represents the ratio between the Monte Carlo MSE obtained with an estimated optimal order and the Monte Carlo MSE obtained under the knowledge of the optimal order.

| **Sample size** | **Beginning of the interval** | | | | | | |
|---|---|---|---|---|---|---|---|
| | $M = \frac{T}{2}$ | $M = \frac{T}{4}$ | $M = \frac{T}{8}$ | $M = \frac{T}{16}$ | $M = \frac{T}{32}$ | $M = \frac{T}{64}$ | $M = \frac{T}{128}$ |
| $T = 50$ | 1.5393 | 1.2905 | 1.1381 | 1.1087 | 1.1037 | – | – |
| $T = 100$ | 1.2669 | 1.1367 | 1.0516 | 1.0439 | 1.0349 | 1.0335 | 1.0406 |
| $T = 150$ | 1.0735 | 1.0313 | 1.0515 | 1.0396 | 1.0352 | 1.0312 | 1.0248 |
| $T = 200$ | 1.2041 | 1.0427 | 1.0418 | 1.0484 | 1.0372 | 1.0302 | 1.0311 |
| $T = 250$ | 1.1768 | 1.0542 | 1.0395 | 1.0042 | 1.0291 | 1.0265 | 1.0346 |
| $T = 300$ | 1.1434 | 1.0780 | 1.0493 | 1.0287 | 0.9921 | 1.0036 | 1.0100 |
| $T = 350$ | 1.2108 | 1.1149 | 1.0573 | 1.0427 | 1.0158 | 1.0065 | 1.0064 |
| $T = 400$ | 1.2251 | 1.1459 | 1.1207 | 1.0750 | 1.0553 | 1.0178 | 1.0008 |
| $T = 450$ | 1.2907 | 1.1332 | 1.0905 | 1.0638 | 1.0382 | 1.0168 | 1.0141 |
| $T = 500$ | 1.2733 | 1.1512 | 1.0736 | 1.0760 | 1.0441 | 1.0327 | 1.0216 |

Table 2.1: Risk Ratios

Table (2.1) shows the values of the index (2.22) for different $M$ and different sample sizes [1] . For falling $M$ one can see a clear convergence toward 1, which means that the risk due to the knowledge of the optimal order and the risk which arises from the estimation of the order eventually tend to coincide.

One has to bare in mind that if the value of $M$ is too small the optimal order is underestimated. Nonetheless the values of the risk are very small even for relatively large values of $M$ like: $M = \frac{T}{2}$ and $M = \frac{T}{4}$; so that for $T \leq 200$, any $M$ between $\frac{T}{8}$ and $\frac{T}{2}$ can be a good choice. For $T > 200$ smaller values of $M$ can also be chosen.

The analysis of the simulation results leads to positive conclusions: the algorithm works well for AR processes, the optimal order is on the average detected, and the loss efficiency which arises from the estimation order (compared with the case when the optimal order is known) is quite small. We do not observe a big sensitivity to varying the values of $M$. Anyway the risk of underfitting the optimal order cannot be excluded, but in practical application this risk can be reduced by comparing the results for different values of $M$.

[1] Note that for the pair ($T = 300$, $M = \frac{T}{32}$) the value of the index is smaller than one, this does not certainly mean that the estimated order is more efficient than the optimal order, actually this value is due to the fact that 100 realisation are not really enough to perform a perfect Monte Carlo simulation, nevertheless the general interpretation of the results is not undermined by that fact.

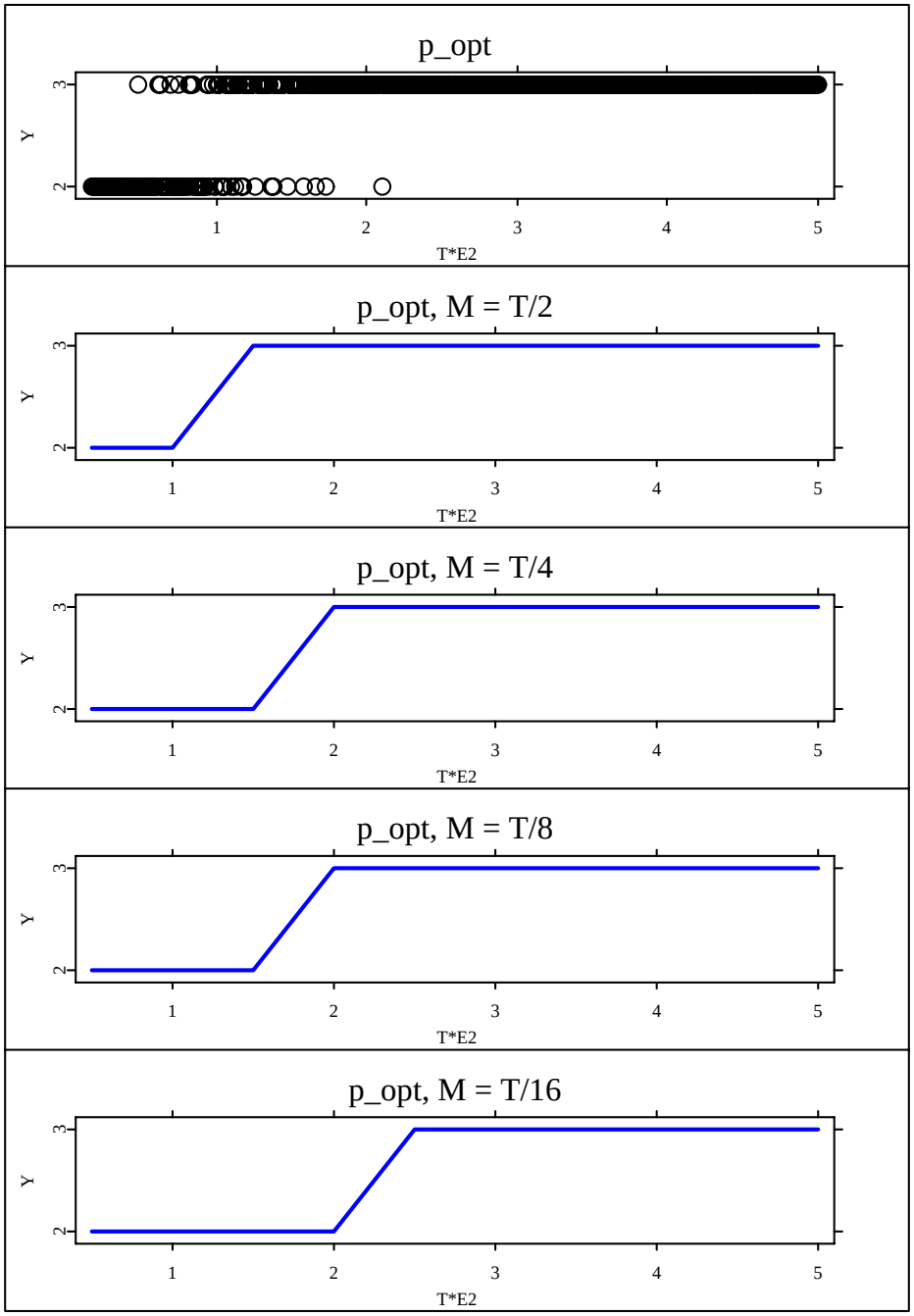


Figure 2.3: The values of $p^{opt}$ for the process (2.19) computed by Monte Carlo according to formulae (2.20) the first, and (2.21) the last four

# Chapter 3

# State space models

In this chapter we introduce the class of the state space models, which can be regarded as a generalisation of the models considered so far. In particular we analyse two examples of such models, which are very similar to two popular models of financial econometrics, namely the EGARCH model and the stochastic volatility model (SV).

The Kalman filter, an extremely useful tool in the analysis of state space models, is briefly introduced along with the problem of adaptive parameter identification. A nice solution to this problem is offered by the extended Kalman filter (EKF), which provides at the same time an estimate of the hidden process and of the parameters. We obtain once again a series of estimates which are only based on the past observations and therefore we can evaluate the sum of the squared forecasting residuals and select among different models.

## 3.1 State space models

State space models are essentially characterised by the fact that an **observed process** depends on an **unobserved, hidden, state process**. Such models are very common in engineering and physics (Elliot et al. (1995) and Chui & Chen (1998)), but they seem to enjoy less popularity in econometrics, although many of the models which are commonly used such as AR, ARMA, and their multivariate homologue VAR, VARMA, Co-integrated and structural time series models can be put in state space form (Lütkepohl (1993, Chapter 4), Harvey (1989) and Aoki (1990)). Actually the shortage of data represents sometimes a constraint for very complex modelling, but in the recent years the increasing availability of high-frequency financial data has created much favourable conditions for the development of state space models at least in financial econometrics. The work of Singer (1998) shows many applications of such models and filtering algorithms for problems of empirical finance.

The **general linear state space model** is a system of two equations which are called the measurement and the state equation, respectively.

$$\begin{cases} \mathbf{y}_t &= A_t\mathbf{z}_t + B_t\mathbf{x}_t + \mathbf{v}_t \\ \mathbf{z}_t &= C_t\mathbf{z}_{t-1} + D_t\mathbf{x}_t + \mathbf{w}_t, \end{cases} \tag{3.1}$$

where:

- $\{\mathbf{y}_t\}_{t=1}^T$ is the observable output sequence,
- $\{\mathbf{z}_t\}_{t=1}^T$ is the hidden state process sequence,
- $\{\mathbf{x}_t\}_{t=1}^T$ is the observable, or deterministic input sequence which may also include lagged values of $\mathbf{y}_t$ ,
- $\{\mathbf{v}_t\}_{t=1}^T$ is the observation noise sequence,

- $\{\mathbf{w}_t\}_{t=1}^T$ is the system noise sequence,
- $A_t$, $B_t$, $C_t$, $D_t$ are the system matrices, which can be time varying and dependent on lagged values of $\mathbf{y}_t$, but they are seen as deterministic and fixed at time $t$
- $\mathrm{E}\,\mathbf{v}_t\mathbf{v}_t^\top = Q_t$ and $\mathrm{E}\,\mathbf{w}_t\mathbf{w}_t^\top = R_t$ are the covariances of the errors; just like the system matrices they don't have to be time homogeneous.

The main problem in the statistical analysis of state space models is the estimation of the state process $\{\mathbf{z}_t\}_{t=1}^T$. Let $\hat{\mathbf{z}}_{t|j}$ be the estimator of $\mathbf{z}_t$ given the observations up to time $j$. We can distinguish three cases: if $j = t$ we have the **filtered** estimate $\hat{\mathbf{z}}_{t|t}$, if $j < t$ we have the **predicted** estimate $\hat{\mathbf{z}}_{t|j}$ and if $j > t$ we have the **smoothed** estimate $\hat{\mathbf{z}}_{t|j}$ of $\mathbf{z}_t$.

As in the case of the autoregressive processes our aim is to compare the sum of the squared forecasting errors and for that reason we are going to focus on the filtering and prediction problem, the smoothing problem being for our purposes not so interesting. Furthermore, we are not going to analyse these problems in full generality but only with the help of two simple examples which suit very well the subsequent study of financial time series. Generalisations to more complex model are possible and, at least in theory, relatively straightforward. Nevertheless one must bare in mind that the need of a parsimonious parametrisation doesn't allow the complexity of the models to grow at any desired extent.

Now we consider two simple examples of state space models which we are going to

analyse later in detail. The first example is represented by the **ARMA(1,1)** model:

$$y_t = \nu + \gamma y_{t-1} + v_t + \phi v_{t-1}; \tag{3.2}$$

a possible state space representation takes the form:

$$\begin{cases} y_t &= z_t + v_t \\ z_t &= \nu + \alpha z_{t-1} + \beta y_{t-1}, \end{cases} \tag{3.3}$$

where the state $z_t := \nu + \gamma y_{t-1} + \phi v_{t-1}$ is a deterministic process of the past observations, and $\alpha = -\phi$ and $\beta = \gamma + \phi$.

The second example of state space model is a first order autoregressive process (state) which is observed with an error, **AR(1) plus noise**:

$$\begin{cases} y_t &= z_t + v_t \\ z_t &= \nu + \theta z_{t-1} + w_t. \end{cases} \tag{3.4}$$

This is a much more typical example of a state space model, because the state vector follows an autonomous stochastic process and we have two white noise sequences.

## 3.2 The Kalman Filter

Consider the system (3.1). If the parameter of the system matrices and the covariances of the errors are known, we can estimate the state process through the Kalman filter, a powerful algorithm attributable to R. E. Kalman (1960), which solves the recursive estimation problem for discrete dynamical systems.

Suppose that the errors $\mathbf{w}_t$ and $\mathbf{v}_t$ are uncorrelated and that some prior information is available so that one can set the following initial conditions: $\hat{\mathbf{z}}_{0|0}$ the estimate of $\mathbf{z}_0$ and $P_{0|0} = \mathrm{E}(\mathbf{z}_0 - \hat{\mathbf{z}}_{0|0})(\mathbf{z}_0 - \hat{\mathbf{z}}_{0|0})^\top$, the initial MSE matrix. Suppose furthermore

that the initial state $\mathbf{z}_0$ is uncorrelated with any of the future errors. Then the **best linear minimum variance estimator** $\hat{\mathbf{z}}_{t|t}$ may be generated recursively by:

$$\begin{cases} P_{t|t-1} &= C_t^\top P_{t-1|t-1} C_t + R_t \\ G_t &= P_{t|t-1} A_t^\top (A_t P_{t|t-1} A_t^\top + Q_t)^{-1} \\ P_{t|t} &= (\mathrm{I} - G_t A_t) P_{t|t-1} \\ \hat{\mathbf{z}}_{t|t-1} &= C_t \hat{\mathbf{z}}_{t-1|t-1} + D_t \mathbf{x}_t \\ \hat{\mathbf{z}}_{t|t} &= \hat{\mathbf{z}}_{t|t-1} + G_t(\mathbf{y}_t - B_t \mathbf{x}_t - A_t \hat{\mathbf{z}}_{t|t-1}), \end{cases} \tag{3.5}$$

where by the best linear minimum variance property it is implied that:

- $\hat{\mathbf{z}}_{t|t}$ is a linear combination of the past and present values of $\mathbf{y}_t$, and
- $\mathrm{E}(\hat{\mathbf{z}}_{t|t} - \mathbf{z}_t)^2$ is minimal with respect to any other linear combination.

It is worth noting that the above results do not require any particular distributional assumption and they are also valid for nonstationary systems. The Kalman filter is optimal if normality holds, otherwise it remains the best among all linear estimators.

### 3.2.1 The extended Kalman filter

Unfortunately the linear Kalman filter cannot be directly implemented if some of the parameters of the process are unknown. The usual solution to this problem consists in estimating first the unknown parameters and run in a second stage the Kalman filter (Harvey 1989), but this approach is not good for our purpose. We want to consider the forecasting performance of a model and for that reason we cannot first estimate the parameters with the whole data set and then make in-sample forecast. We have to find an algorithm which at the same time updates the estimate of the state vector and of the parameters. An acceptable solution to this problem is given by the **extended Kalman filter** (EKF).

Consider the following nonlinear system:

$$\begin{cases} \mathbf{y}_t &= f_t(\mathbf{z}_t) + \mathbf{v}_t \\ \mathbf{z}_t &= g_t(\mathbf{z}_{t-1}) + \mathbf{w}_t, \end{cases} \tag{3.6}$$

where $f_t(\cdot)$ and $g_t(\cdot)$ are vector valued functions and the subscript $t$ indicates furthermore that they can be time varying.

Obtaining an optimal filter for a model of this kind is, in general, not possible; however an approximate filter can be obtained by linearising an then applying a modification of the usual Kalman filter. Suppose that $f_t(\cdot)$ and $g_t(\cdot)$ are sufficiently smooth so that they can be approximated by a first order Taylor expansion:

$$\begin{cases} f_t(\mathbf{z}_t) &\approx f_t(\hat{\mathbf{z}}_{t|t-1}) + \left[\frac{\partial f_t(\hat{\mathbf{z}}_{t|t-1})}{\partial \mathbf{z}_t}\right](\mathbf{z}_t - \hat{\mathbf{z}}_{t|t-1}) \\ g_t(\mathbf{z}_t) &\approx g_t(\hat{\mathbf{z}}_{t|t}) + \left[\frac{\partial g_t(\hat{\mathbf{z}}_{t|t})}{\partial \mathbf{z}_t}\right](\mathbf{z}_t - \hat{\mathbf{z}}_{t|t}), \end{cases}$$

where:

$$\left[\frac{\partial f_t(\hat{\mathbf{z}}_{t|t-1})}{\partial \mathbf{z}_t}\right] := \left[\frac{\partial f_t(\mathbf{z}_t)}{\partial \mathbf{z}_t}\right]_{\mathbf{z}_t=\hat{\mathbf{z}}_{t|t-1}} \quad \text{and} \quad \left[\frac{\partial g_t(\hat{\mathbf{z}}_{t|t})}{\partial \mathbf{z}_t}\right] := \left[\frac{\partial g_t(\mathbf{z}_t)}{\partial \mathbf{z}_t}\right]_{\mathbf{z}_t=\hat{\mathbf{z}}_{t|t}}$$

The linear approximation of model (3.6) follows straightforwardly:

$$\begin{cases} \mathbf{y}_t &\approx \left[\frac{\partial f_t(\hat{\mathbf{z}}_{t|t-1})}{\partial \mathbf{z}_t}\right]\mathbf{z}_t + \left\{f_t(\hat{\mathbf{z}}_{t|t-1}) - \left[\frac{\partial f_t(\hat{\mathbf{z}}_{t|t-1})}{\partial \mathbf{z}_t}\right]\hat{\mathbf{z}}_{t|t-1}\right\} + \mathbf{v}_t \\ \mathbf{z}_t &\approx \left[\frac{\partial g_{t-1}(\hat{\mathbf{z}}_{t-1|t-1})}{\partial \mathbf{z}_{t-1}}\right]\mathbf{z}_{t-1} + \left\{g_{t-1}(\hat{\mathbf{z}}_{t-1|t-1}) - \left[\frac{\partial g_t(\hat{\mathbf{z}}_{t-1|t-1})}{\partial \mathbf{z}_{t-1}}\right]\hat{\mathbf{z}}_{t-1|t-1}\right\} + \mathbf{w}_t, \end{cases}$$

where the Jacobi matrices and the expressions in braces are known and deterministic at each time $t$. Now we can easily apply the Kalman filter recursion described in (3.5), with the exception that the predictors for $\mathbf{z}_t$ and $\mathbf{y}_t$ given $\hat{\mathbf{z}}_{t-1|t-1}$ are $\hat{\mathbf{z}}_{t|t-1} = g_{t-1}(\hat{\mathbf{z}}_{t-1|t-1})$ and $\hat{\mathbf{y}}_{t|t-1} = f_t(\hat{\mathbf{z}}_{t|t-1})$ respectively.

The EKF algorithm can be summarised by the following equations:

$$\begin{cases} P_{t|t-1} &= \left[\frac{\partial g_t(\hat{\mathbf{z}}_{t-1|t-1})}{\partial \mathbf{z}_{t-1}}\right] P_{t-1|t-1} \left[\frac{\partial g_t(\hat{\mathbf{z}}_{t-1|t-1})}{\partial \mathbf{z}_{t-1}}\right]^\top + R_t \\ \hat{\mathbf{z}}_{t|t-1} &= g_t(\hat{\mathbf{z}}_{t-1|t-1}) \\ G_t &= P_{t|t-1} \left[\frac{\partial f_t(\hat{\mathbf{z}}_{t|t-1})}{\partial \mathbf{z}_t}\right]^\top \\ & \left[\left[\frac{\partial f_t(\hat{\mathbf{z}}_{t|t-1})}{\partial \mathbf{z}_t}\right] P_{t|t-1} \left[\frac{\partial f_t(\hat{\mathbf{z}}_{t|t-1})}{\partial \mathbf{z}_t}\right]^\top + Q_t\right]^{-1} \\ P_{t|t} &= \left[\mathrm{I} - G_t \left[\frac{\partial f_t(\hat{\mathbf{z}}_{t|t-1})}{\partial \mathbf{z}_t}\right]\right] P_{t|t-1} \\ \hat{\mathbf{z}}_{t|t} &= \hat{\mathbf{z}}_{t|t-1} + G_t(\mathbf{y}_t - f_t(\hat{\mathbf{z}}_{t|t-1})) \end{cases} \tag{3.7}$$

### 3.2.2 Parameter identification

The EKF is also very useful if some of the parameters in the system matrices of the general linear model (3.1) are unknown. In that case, these unknown parameters are treated as stochastic processes. The model becomes nonlinear and the EKF can be applied to estimate the state process and the parameters (Chui & Chen (1998, Chapter 8) and Singer (1998)).

Let's consider once again our simple examples of state space models: the ARMA(1,1) (3.3) and the AR(1) plus error (3.4). Suppose that apart form the observed process $\{\mathbf{y}_t\}_{t=1}^T$, we know only the variances of the two noise sequences $\{v_t\}_{t=1}^T$ and $\{w_t\}_{t=1}^T$; the parameters $\nu$, $\alpha$, $\beta$, and $\theta$ are unknown and therefore one has to consider them has stochastic processes.

#### Estimation of the ARMA(1,1)

Consider the ARMA(1,1) model (3.3), and suppose that the parameters are unknown. Now make the technical assumption that the parameters follow a stochastic process: a random walk. In this case the ARMA(1,1) model may be written in the following

state space form:

$$\begin{cases} y_t & = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_t \\ \alpha_t \\ \beta_t \\ \nu_t \end{bmatrix} + v_t \\ \begin{bmatrix} z_t \\ \alpha_t \\ \beta_t \\ \nu_t \end{bmatrix} & = \begin{bmatrix} \nu_t + \alpha_{t-1} z_{t-1} + \beta_{t-1} y_{t-1} \\ \alpha_{t-1} \\ \beta_{t-1} \\ \nu_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ e_{1t} \\ e_{2t} \\ e_{3t} \end{bmatrix}, \end{cases} \tag{3.8}$$

The above system is nonlinear because of the term $\alpha_{t-1} z_{t-1}$, so that for the estimation of $z_t$, $\alpha_t$ and $\beta_t$ one can apply the EKF (3.7). As usual one has to set some initial values for $\hat{z}_{0|0}$, $\hat{\alpha}_{0|0}$, $\hat{\beta}_{0|0}$, $\hat{\nu}_{0|0}$ and $P_{0|0}$. Then the recursive estimation of the state

process according to (3.7) is implemented as follows:

$$
\begin{cases}
\begin{bmatrix} \hat{z}_{t|t-1} \\ \hat{\alpha}_{t|t-1} \\ \hat{\beta}_{t|t-1} \\ \hat{\nu}_{t|t-1} \end{bmatrix} = \begin{bmatrix} \hat{\nu}_{t-1|t-1} + \hat{\alpha}_{t-1|t-1}\hat{z}_{t-1|t-1} + \hat{\beta}_{t-1|t-1}y_{t-1|t-1} \\ \hat{\alpha}_{t-1|t-1} \\ \hat{\beta}_{t-1|t-1} \\ \hat{\nu}_{t-1|t-1} \end{bmatrix} \\
P_{t|t-1} = \begin{bmatrix} \hat{\alpha}_{t-1|t-1} & \hat{z}_{t-1|t-1} & y_{t-1} & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} P_{t-1|t-1} \\
\qquad \begin{bmatrix} \hat{\alpha}_{t-1|t-1} & 0 & 0 & 0 \\ \hat{z}_{t-1|t-1} & 1 & 0 & 0 \\ y_{t-1} & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \mathrm{Var}(e_{1t}) & 0 & 0 \\ 0 & 0 & \mathrm{Var}(e_{2t}) & 0 \\ 0 & 0 & 0 & \mathrm{Var}(e_{3t}) \end{bmatrix} \\
G_t = P_{t|t-1} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \left[ \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} P_{t|t-1} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mathrm{Var}(v_t) \right]^{-1} \\
P_{t|t} = \left( \mathrm{I} - G_t \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \right) P_{t|t-1} \\
\begin{bmatrix} \hat{z}_{t|t} \\ \hat{\alpha}_{t|t} \\ \hat{\beta}_{t|t} \end{bmatrix} = \begin{bmatrix} \hat{z}_{t|t-1} \\ \hat{\alpha}_{t|t-1} \\ \hat{\beta}_{t|t-1} \end{bmatrix} + G_t \left( y_t - \hat{\nu}_{t-1|t-1} - \hat{\alpha}_{t-1|t-1}\hat{z}_{t-1|t-1} - \hat{\beta}_{t-1|t-1}y_{t-1|t-1} \right)
\end{cases}
\tag{3.9}
$$

### Estimation of the AR(1) plus noise

The estimation technique for the AR(1) plus noise can be derived in an analogous way. First one has to consider the nonlinear system:

$$\begin{cases} y_t = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_t \\ \theta_t \\ \nu_t \end{bmatrix} + v_t \\ \begin{bmatrix} z_t \\ \theta_t \\ \nu_t \end{bmatrix} = \begin{bmatrix} \nu_{t-1} + \theta_{t-1} z_{t-1} \\ \theta_{t-1} \\ \nu_{t-1} \end{bmatrix} + \begin{bmatrix} w_t \\ e_{1t} \\ e_{2t} \end{bmatrix}, \end{cases} \tag{3.10}$$

where the nonlinearity arises because of the term $\theta_{t-1} z_{t-1}$. Then one has to set some starting values for $\hat{z}_{0|0}$, $\hat{\theta}_{0|0}$, $\hat{\nu}_{0|0}$ and $P_{0|0}$, and finally one can run the EKF:

$$\begin{cases} \begin{bmatrix} \hat{z}_{t|t-1} \\ \hat{\theta}_{t|t-1} \\ \hat{\nu}_{t|t-1} \end{bmatrix} = \begin{bmatrix} \hat{\nu}_{t-1|t-1} + \hat{\theta}_{t-1|t-1} \hat{z}_{t-1|t-1} \\ \hat{\theta}_{t-1|t-1} \\ \hat{\nu}_{t-1|t-1} \end{bmatrix} \\ P_{t|t-1} = \begin{bmatrix} \hat{\theta}_{t-1|t-1} & \hat{z}_{t-1|t-1} & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} P_{t-1|t-1} \\ \qquad \begin{bmatrix} \hat{\theta}_{t-1|t-1} & 0 & 0 \\ \hat{z}_{t-1|t-1} & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} + \begin{bmatrix} \mathrm{Var}(w_t) & 0 & 0 \\ 0 & \mathrm{Var}(e_{1t}) & 0 \\ 0 & 0 & \mathrm{Var}(e_{2t}) \end{bmatrix} \\ G_t = P_{t|t-1} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \left[ \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} P_{t|t-1} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \mathrm{Var}(v_t) \right]^{-1} \\ P_{t|t} = \left( \mathrm{I} - G_t \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \right) P_{t|t-1} \\ \begin{bmatrix} \hat{z}_{t|t} \\ \hat{\theta}_{t|t} \\ \hat{\nu}_{t|t} \end{bmatrix} = \begin{bmatrix} \hat{z}_{t|t-1} \\ \hat{\theta}_{t|t-1} \\ \hat{\nu}_{t|t-1} \end{bmatrix} + G_t \left( y_t - \hat{\nu}_{t-1|t-1} - \hat{\theta}_{t-1|t-1} \hat{z}_{t-1|t-1} \right) \end{cases} \tag{3.11}$$

## 3.3 Optimising the filter

The EKF is a very flexible tool, because it allows a **real time, recursive, explicit estimation of the state process and of the parameter values** of models like (3.3) and (3.4) and therefore recursive forecasts of the observed process, furthermore simulation results in Chui & Chen (1998) show that it can take in account parameter changes.

The drawback is that one has to choose many starting values: for the state process $\hat{z}_{0|0}$, for the parameters $\hat{\alpha}_{0|0}$ and $\hat{\beta}_{0|0}$, or $\hat{\theta}_{0|0}$ and for $P_{0|0}$. Furthermore a value for the variances of the innovations of the parameters $e_{it}$ must be chosen; and finally the variances of innovation $v_t$ and $w_t$ are supposed to be known.

Usually the starting values are chosen according to some prior information. They may express the unconditional expectation of the process, if only a very limited amount of prior information is available. Otherwise they are derived from the past experience of the researcher; the initial MSE matrix $P_{0|0}$ expresses the degree of uncertainty about the starting values: the larger the MSE, the greater the uncertainty about the parameters.

The choice of the variances of the innovation of the parameters $\mathrm{Var}(e_{it})$, is of essential importance for the EKF. In fact these values act as a weight which evaluates the deviation of the predicted estimate $\hat{y}_{t|t-1}$ from the actual realization of the observed process $y_t$. Therefore, if those variances are too large the parameter estimation is too noisy, but if they are too small the recursive estimator converges too slowly toward the true value of the parameter.

In conclusion, if we fix the initial conditions in a wrong way the filtering algorithm works badly and produces very poor forecasts. For that reason, we must find a way of selecting good starting values.

A possible solution to this problem is to consider the initial conditions as parameters that have to be estimated. For that purpose, we can apply the theory that we developed in Chapter 1: we compare the forecasting ability of the estimates under different sets of initial conditions. For any value of the observed process $y_t$ we have a predicted estimate which is based only on the observation up to time $t-1$ and therefore we can analyse the squared forecasting residuals, namely for the ARMA(1,1) model:

$$\sum_{t=M}^{T} \left( y_t - \hat{\nu}_{t-1|t-1} - \hat{\alpha}_{t-1|t-1}\hat{z}_{t-1|t-1} - \hat{\beta}_{t-1|t-1} y_{t-1} \right)^2 ,$$

and for AR(1) plus noise model:

$$\sum_{t=M}^{T} \left( y_t - \hat{\nu}_{t-1|t-1} - \hat{\theta}_{t-1|t-1}\hat{z}_{t-1|t-1} \right)^2 .$$

These two expressions are specific examples of the loss function in (1.8), whose minimisation represents our model selection criterion, and therefore we can identify the optimal set of starting staring values which leads to the best forecasting performance.

Obviously if we face a real data set we don't know the underlying true model, but the analysis of the sum of the squared forecasting residuals provides a useful tool for selecting, not only among different orders or among different sets of starting values but also among different models. Indeed, we are now able to evaluate which model produces the best forecasting performance.

# Chapter 4

# Volatility models

This chapter is concerned with the question of model selection for volatility models. The Strategy of optimal model selection is applied to three volatility, models, exponential-ARCH (EARCH), EGARCH and stochastic volatility SV, with which the volatility of daily financial time series for different sample sizes is predicted.

## 4.1 Stylised facts of financial time series

Financial price series $\{P_t\}_{t=1}^{T}$ such as stocks, exchange rates or price indices are generally modelled both in financial theory and in financial econometrics as a martingale process (Gouriéroux 1997, pagg. 83-90), which means that:

$$\mathrm{E}(P_{t+1}|P_t) = P_t, \qquad \forall\ t. \tag{4.1}$$

The above relationship is connected to the hypothesis of efficiency of the financial markets and implies that the best forecast of the future price at date $t$ is the current price $P_t$.

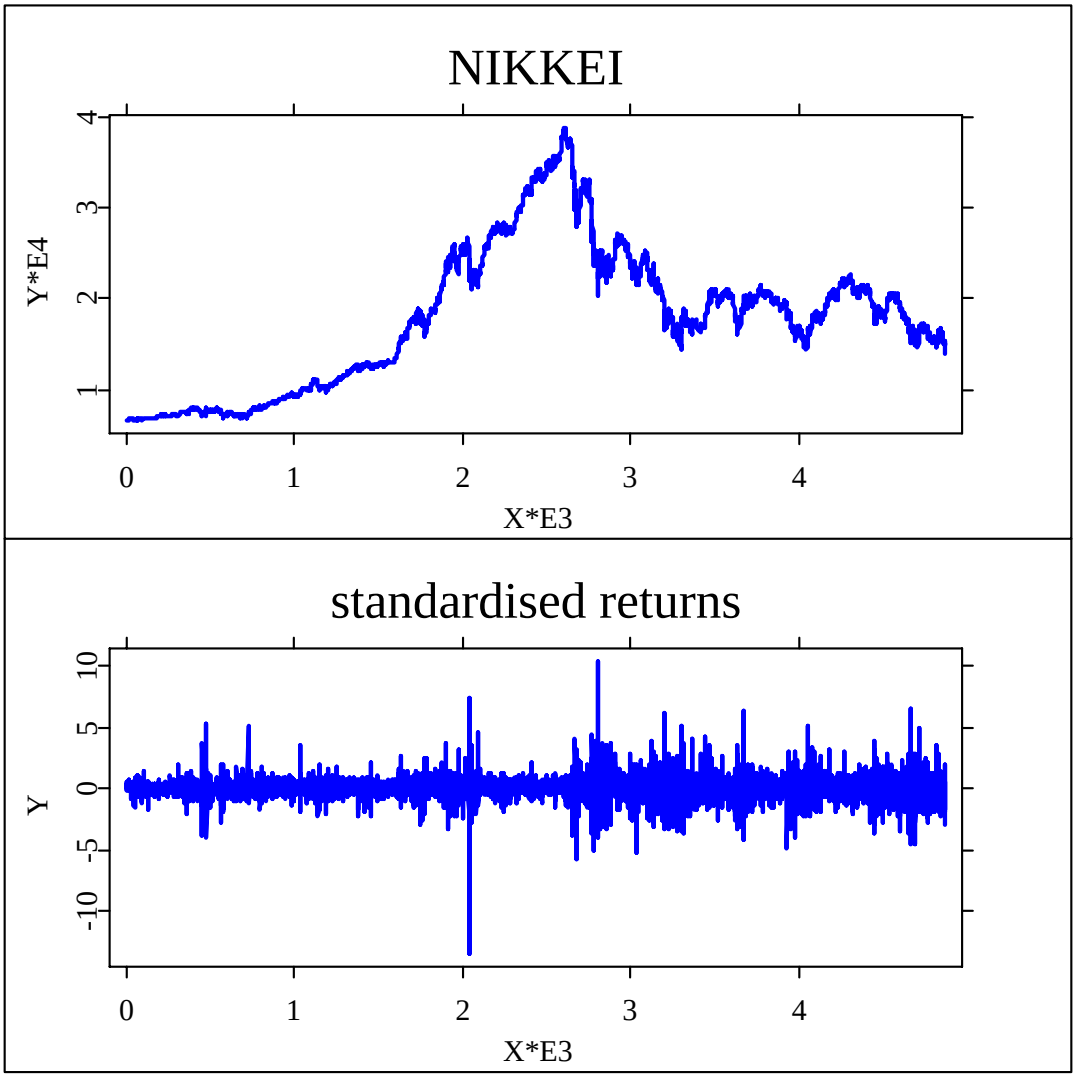


Figure 4.1: NIKKEI stock index (upper plot) and the standardised returns (lower plot).

Most of the empirical work on financial time series does not focus on the prices, but on their growth rates, i.e. the returns:

$$r_t = \ln P_t - \ln P_{t-1}, \tag{4.2}$$

which, under the hypothesis of efficiency, are modelled by a zero mean uncorrelated process, a white noise.

Although uncorrelated, the returns show dependency in higher moments. In particular the variance of $r_t$ does not appear to be constant. Indeed, one can observe the so called clustering effect : large changes in the return tend to be followed by large changes of either sign and vice versa. Furthermore the distribution of the returns shows a very high kurtosis if compared with the standard normal distribution.

The daily NIKKEI stock index (from 01.01.1980 to 30.12.1988) is considered as a graphical example. For comparison with the standard normal distribution the standardised residuals, i.e. centred and divided by the sample standard deviation, are plotted. In the lower plot of Figure 4.1 the clustering effect may be appreciated, while Figure 4.3 shows the empirical density of the standardised returns against the empirical density of the realisation of a standard normal random variable of the same sample size (4868). It can easily be seen that the density of the returns is much more fat tailed (leptokurtic) than the standard normal density.

The observation of the sample autocorrelations (Hafner 1998) gives a further insight in the stylised fact of the financial time series and underlines the fact that the financial returns are not independent. In particular the autocorrelation function of a transformation of the returns from $\mathbb{R}$ to $\mathbb{R}_+$, such as the absolute value of the returns,

the squared, or the log-squared returns, appears to be highly significant even for very high lags (see Figure 4.2). The behaviour of these autocorrelation functions is deeply connected with the volatility clustering effect and it leads to the conclusion that an heteroskedastic process is required to model the returns of financial time series.

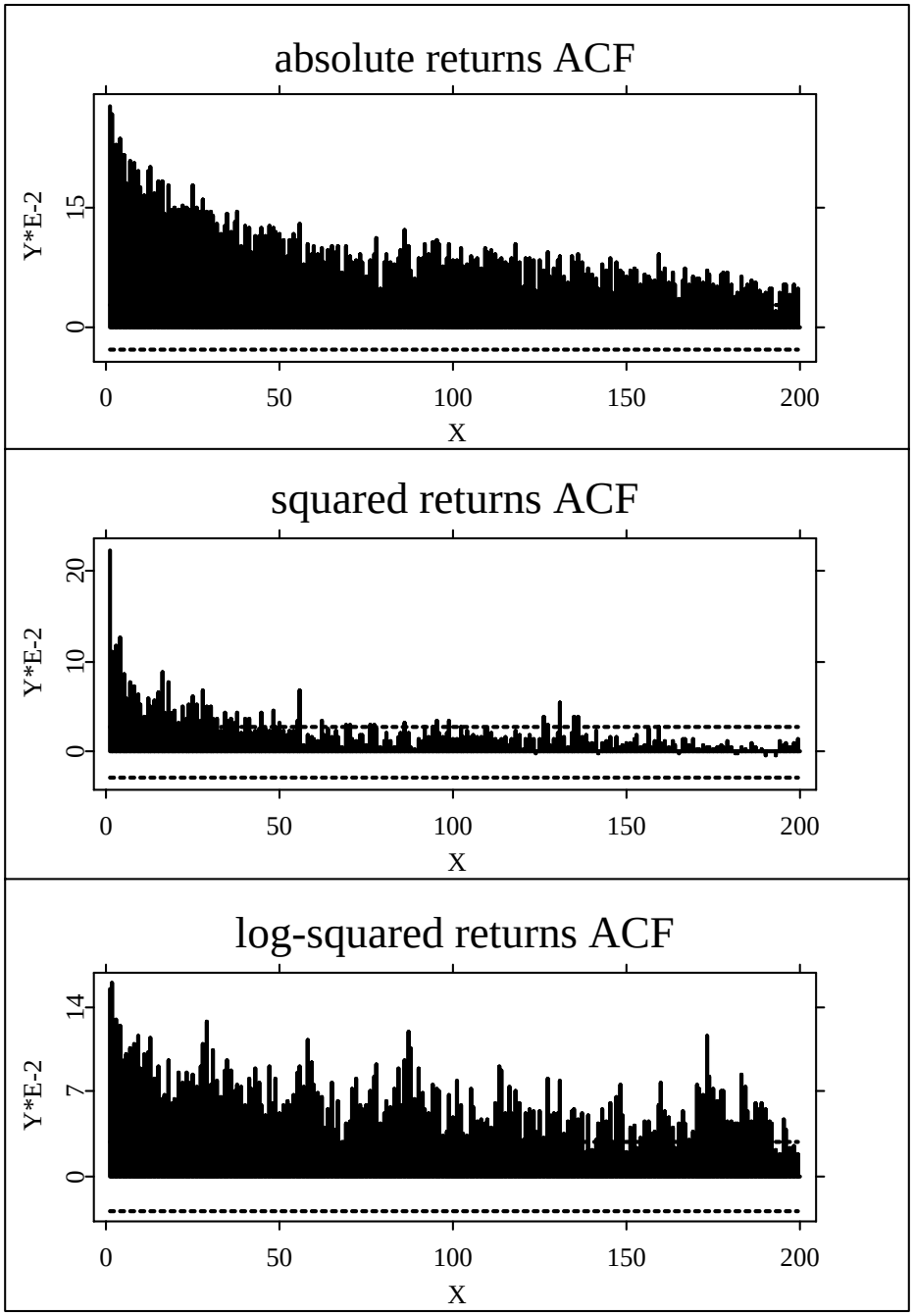


Figure 4.2: Autocorrelations of the absolute values of the returns, of the squared and of the log-squared returns of the NIKKEI stock index.

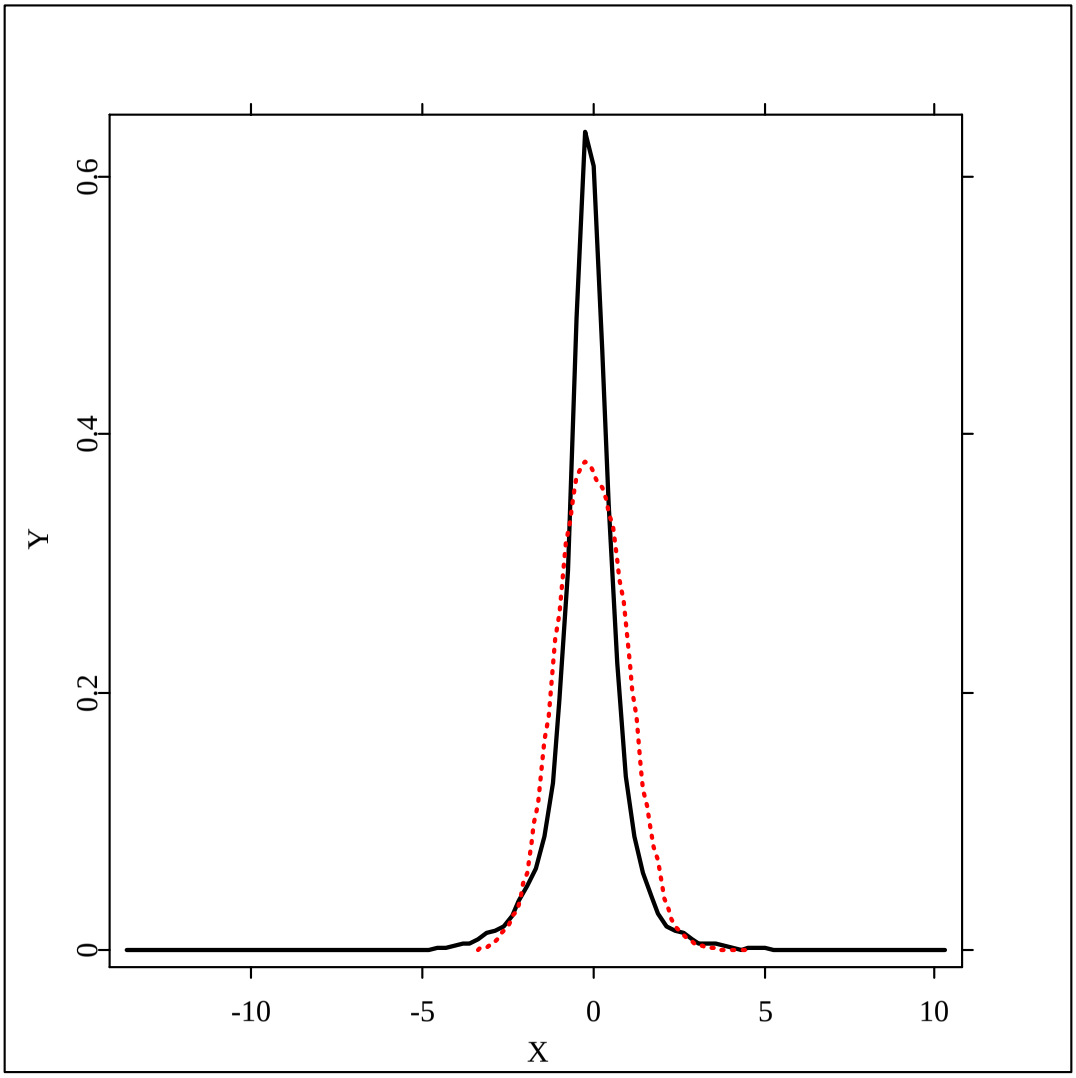


Figure 4.3: Empirical densities of a standard normal sample (dotted) and of the standardised returns of the NIKKEI stock index (straight line).

## 4.2 Volatility models

Many processes have been proposed which can reproduce the characteristics of volatility clustering and leptokurtosis of the financial time series (see Bollerslev et al. (1992) and Engle (1995b) for a collection of the most famous articles about that topic and Gouriéroux (1997) for a focus on the implication for the financial theory and praxis). In particular, the most famous among them are: the ARCH model (autoregressive conditional heteroscedasticity), the GARCH model (generalised ARCH), the EGARCH model (exponential GARCH) and the stochastic volatility model.

In this study we want to analyse the forecasting performances of three volatility models and select the one which shows the best forecasting ability. For the application the model selection algorithm developed in the previous chapters models with additive errors are preferable. Therefore, we focus on the log-volatility. Consider the following representation of the returns:

$$r_t = \sigma_t \varepsilon_t, \tag{4.3}$$

where $\sigma_t$ is a positive valued process which represents the standard deviation at time $t$ and $\varepsilon_t \sim N(0,1)$ is a standard Gaussian white noise. Furthermore, $\sigma_t$ is supposed to be independent from the present and future errors.

The distribution assumption is quite strong, but is required to identify the volatility, because expected value and variance of the log-squared innovations: $\ln \varepsilon_t^2$, are needed. Under the normality assumption these quantities are:

$$\mathrm{E} \ln \varepsilon_t^2 \approx -1.27 \qquad \text{and} \qquad \mathrm{Var} \ln \varepsilon_t^2 \approx \frac{\pi^2}{2}.$$

Nevertheless this assumption is not inconsistent with the leptokurtosis of the return,

because it implies only conditional normality.

After squaring and taking the natural logarithms we obtain the following linear model:

$$\ln r_t^2 = \ln \sigma_t^2 + \ln \varepsilon_t^2,$$

or equivalently:

$$y_t = h_t + v_t \tag{4.4}$$

where

$$y_t := \ln r_t^2, \qquad h_t := \ln \sigma_t^2 + \mathrm{E} \ln \varepsilon_t^2 \qquad \text{and} \qquad v_t := \ln \varepsilon_t^2 - \mathrm{E} \ln \varepsilon_t^2.$$

The variable $h_t$ represents, up to a constant, the volatility process that we want to predict, and equation (4.4) can be interpreted as the observation equation of a state space model. The true state process is unknown and we consider three possible modelling assumptions whose characteristics are consistent with the stylised facts of the financial returns: volatility clustering and leptokurtosis.

### 4.2.1 Exponential-ARCH

The first model which is analysed is an exponential version of the ARCH model: EARCH, where the volatility is a deterministic processes of the past realisations of the returns.

$$h_t = \nu + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} \tag{4.5}$$

Equation (4.5) is very similar to the original ARCH modelling philosophy, where the focus is set directly on the variance and on the squared returns:

$$\sigma_t^2 = \nu + \beta_1 r_{t-1}^2 + \ldots + \beta_p r_{t-p}^2.$$

Both perspectives have their own advantages and disadvantages, in particular the ARCH model does not need particular distribution assumptions, but it requires constrained estimation techniques to guarantee the positiveness of the parameters and of the volatility process.

ARCH represents the first attempt of modelling the volatility dynamics of financial time series (Engle 1995a). This kind of modelling is in general not very parsimonious because it usually demands a very high number of lags. Actually, the observation of the autocorrelation function provides some empirical evidence of the fact that the volatility process is smooth and very persistent. Therefore, a modelling strategy which only relies on the past realisations of the returns will require a quite large order. On the other hand, there are some advantages because equations (4.4) and (4.5) can be summarised in an autoregressive process in $y_t$ so that least square may be used for estimation. For the lag selection, the theory developed in Chapter 2 can be easily implemented.

### 4.2.2 EGARCH

A more parsimonious representation can be gained if one considers the generalised version of the ARCH process (Bollerslev 1995), where the actual value of the volatility depends not only on the past realisations of the returns, but also on its own past realisation and therefore it can express much more persistent dynamics. In fact, under stationarity condition a generalised ARCH has an infinite ARCH representation. In particular we consider the following EGARCH(1,1), where equation (4.6) together with equation (4.4) expresses the state space representation of the ARMA(1,1) that

we have seen in equations (3.2) and (3.3):

$$h_t = \nu + \alpha h_{t-1} + \beta y_{t-1}. \tag{4.6}$$

The EGARCH model was first proposed by Nelson (1995) in a slightly more general framework; he assumes for the $\varepsilon_t$ a generalised error distribution, which includes the normal distribution as a special case, and in his model the log-volatility does not depend on the log squared returns but on a function of the errors:

$$g(\varepsilon_t) = a\varepsilon_t + b(|\varepsilon_t| - \mathrm{E}\,|\varepsilon_t|),$$

allowing therefore for asymmetry in the volatility process. The estimation of the parameters is usually performed with maximum likelihood.

In principle, asymmetry could be easily introduced also in our version of the EGARCH model. On the other hand, for our model selection technique we need to perform the estimation at each date $t$ and therefore an estimator which require a numerical optimisation such as maximum likelihood does not seem to be the most suitable choice. Actually, equation (4.6) is not essentially different from the original EGARCH model and has the advantage of allowing the implementation of the extended Kalman filter which was introduced in the previous chapter in equation (3.8) and (3.9).

### 4.2.3 Stochastic volatility (SV)

The last model that we want to consider differs from the ARCH family in a particular aspect, i.e. the volatility follows an autonomous stochastic process:

$$h_t = \nu + \theta h_{t-1} + w_t, \tag{4.7}$$

The stochastic volatility model has become increasingly popular in recent years and it represents the rival model class to the ARCH models. It has the nice feature of being the natural discrete-time version of the continuous-time models which are commonly used in the framework of mathematical finance and stochastic analysis of derivatives. Its drawback consists in the more complicated estimation techniques which are needed to estimate its parameters.

In the first article on this theme Harvey et al. (1995) proposed quasi-maximum likelihood, and in the recent years many publications have appeared, which consider more efficient algorithms, such as: indirect inference (Monfardini 1996), Monte Carlo Markov Chain (Jacquier, Polson & Rossi 1994), simulated maximum likelihood (Danielsson 1994) and simulated method of moments (Gouriéroux 1997, pagg 82-83).

Nevertheless all these methods are not very suitable for our purposes because they require numerical simulations and optimisation techniques and we need to perform the estimation at each date $t$, therefore they would lead to a huge computational effort. For this reason we rely, as for the EGARCH model, on the extended Kalman filter. Our estimation technique may be criticised because Kalman filter type algorithms are optimal only if the errors $v_t$ and $w_t$ are normally distributed, and here it is precisely not the case. Indeed, $v_t = \ln \varepsilon_t^2$ is very far from being normal (for the same reason the quasi-maximum likelihood is quite unpopular), but there are some advantages: the estimator is flexible as far as forecasting is concerned, because at each step it produces an out of sample forecast. It takes into account parameter changes (Chui & Chen 1998, pagg. 124-28). Furthermore because of the rich availability of financial data the question of efficiency loses some of its importance.

## 4.3 Empirical evidence

In this section we compare the forecasting performances of the three modelling strategies, which have been presented above: EARCH($p$), EGARCH and SV. The following data sets are taken into consideration:

- Japanese-Yen/US-Dollar exchange rate, from 02.01.1986-28.8.1998, daily.
- ECU/US-Dollar exchange rate, from 01.01.1980-30.12.1998, daily.
- Standard and Poor 500 composite price index, from 01.01.1980-30.12.1988, daily.
- NIKKEI 225 stock average price index, from 01.01.1980-30.12.1988, daily.
- Volkswagen stock price, from 01.01.1980-30.12.1988, daily.

To enable the comparison of the results, only the first 3000 observations of each data set are taken into consideration for the model selection. Furthermore the NIKKEI data set is used again as a graphical example.

### 4.3.1 The problem of missing observations

The analysis of real data has some specific problems. In particular, as far as the models for the log-volatility are concerned it may happen that two successive prices are equal, so that the return becomes zero, and the relative the log-squared return sample $\{y_t\}_{t=1}^T := \{\ln r_t^2\}_{t=1}^T$ have some missing observations.

In this study the problem is solved in two different ways: for the computation of the autocorrelations, of the LS estimator for the EARCH model and for the determination

of the initial conditions, the missing observations are substituted by the minimum real value attained by the $\{y_t\}_{t=1}^T$, so that we obtain an adjusted sample $\{y_t^*\}_{t=1}^T$. For the estimation of the EGARCH and SV models the missing are treated in the framework of the Kalman Filter (Harvey 1992, pag. 95), which enables to solve the problem by skipping the updating equations. Thus, following the notation of the general model (3.5), if $\mathbf{y}_t$ is not available:

$$\mathbf{z}_{t|t} = \mathbf{z}_{t|t-1} \qquad \text{and} \qquad P_{t|t} = P_{t|t-1}.$$

As far as the model selection criterion is concerned:

$$\sum_{t=M}^{T} \left( y_t - \hat{f}_t(\psi, Y_{t-1}) \right)^2,$$

these summands, where the value $y_t$ is a missing observation, are omitted. In the computation of the forecasts no problems arise for the SV model because:

$$\hat{f}_t(\psi, Y_{t-1}) = \hat{\nu}_{t-1|t-1} + \hat{\theta}_{t-1|t-1}\hat{h}_{t-1|t-1} \qquad \text{SV forecast,}$$

does not require any value of $y_t$; the EGARCH and EARCH models on the other hand contain values of $y_t$:

$$\begin{aligned} \hat{f}_t(\psi, Y_{t-1}) &= \hat{\nu}_{t-1|t-1} + \hat{\alpha}_{t-1|t-1}\hat{h}_{t-1|t-1} + \hat{\beta}_{t-1|t-1}y_t && \text{EGARCH forecast} \\ \hat{f}_t(\psi, Y_{t-1}) &= \hat{\nu}_{t-1|t-1} + \hat{\beta}_{1,t-1|t-1}y_{t-1} + \ldots + \hat{\beta}_{p,t-1|t-1}y_{t-p} && \text{EARCH}(p) \text{ forecast.} \end{aligned}$$

Therefore in the case of a missing observation the value of $y_t$ is substituted by its own forecast.

### 4.3.2 Model selection in practice

The theoretical consideration about the selection of the optimal model, presented in Chapter 1, are now implemented in a practical example. We recall the general definition of the estimator of the optimal model:

$$\hat{\psi} = \arg\inf_{\psi \in \Psi} \sum_{t=M}^{T} \left( y_t - \hat{f}_t(\psi, Y_{t-1}) \right)^2 .$$

In the specific context of the volatility models that we have introduced the set $\Psi$ includes the EARCH($p$) models, with $p$ between $p^{min} = 1$ and $p^{max} = 14$ and the EGARCH and SV models, which in the framework of the extended Kalman filter, require some initial values. These initial values do not differ conceptually from the order of the EARCH($p$) process. In both cases, the selection of the order, or the selection of the initial values completely defines the outcome of the estimation. Therefore in practice we select between the EARCH($p$) models, with $p \in [1, 14]$, and the SV and EGARCH models for different sets of initial values.

For EGARCH and the SV model a grid search is made to select good initial conditions; in particular the values listed in Table 4.1 are common to any data set, while Table 4.2 contains the specific initial conditions. We remark that the analysis of the data has shown that moderate changes in the values of the initial conditions do not lead to substantial changes in the forecasting performance, at least in the long run.

| | |
|---|---|
| $h_{0\|0}$ | $\mathrm{E}\, y_t$ |
| $\mathrm{Var}\, e_{it}$ | $10^{-7}$ |
| $P_{0\|0}$ | $\mathrm{diag}(\mathrm{Var}\, y_t, \mathrm{Var}\, e_{it}, \ldots)$ |
| $\mathrm{Var}\, v_t$ | $\frac{\pi^2}{2}$ |
| $\mathrm{Var}\, w_t$ | 0.05 |

Table 4.1: Common initial conditions

| | | |
|---|---|---|
| Yen/Dollar | EGARCH | $\alpha_{0\|0} = 0.84$ $\beta_{0\|0} = 0.015$ $\nu_{0\|0} = -1.6$ |
| | SV | $\theta_{0\|0} = 0.91$ $\nu_{0\|0} = -1$ |
| ECU/Dollar | EGARCH | $\alpha_{0\|0} = 0.84$ $\beta_{0\|0} = 0.015$ $\nu_{0\|0} = -1.55$ |
| | SV | $\theta_{0\|0} = 0.91$ $\nu_{0\|0} = -1$ |
| Standard and Poor | EGARCH | $\alpha_{0\|0} = 0.8$ $\beta_{0\|0} = 0.019$ $\nu_{0\|0} = -1.9$ |
| | SV | $\theta_{0\|0} = 0.92$ $\nu_{0\|0} = -0.89$ |
| NIKKEI | EGARCH | $\alpha_{0\|0} = 0.8$ $\beta_{0\|0} = 0.019$ $\nu_{0\|0} = -1.8$ |
| | SV | $\theta_{0\|0} = 0.92$ $\nu_{0\|0} = -0.89$ |
| Volkswagen | EGARCH | $\alpha_{0\|0} = 0.77$ $\beta_{0\|0} = 0.02$ $\nu_{0\|0} = -1.76$ |
| | SV | $\theta_{0\|0} = 0.91$ $\nu_{0\|0} = -0.81$ |

Table 4.2: Specific initial conditions

For each data set the value of the sum of the squared forecast errors is computed for three different sample sizes: 500, 1000 and 3000 observations and for three different values of $M$: $\frac{T}{2}$, $\frac{T}{3}$ and $\frac{T}{4}$.

The Tables 4.3, 4.4, 4.5, 4.6 and 4.7 contain the results of the model selection for each data set: the performances of the EARCH($\hat{p}$), with the estimated optimal order, and of the SV and EGARCH models, relative to the initial conditions listed in Tables 4.1 and 4.2, are presented.

The results are among all data sets qualitative very similar and provide good evidence of the fact that the EARCH($p$) modelling strategy is outperformed by the EGARCH and SV models. It remains competitive only for moderate sample sizes ($T = 500$ and $T = 1000$) and only for the Standard and Poor, and for the ECU/US-Dollar exchange rate data sets. For all other data sets the SV and EGARCH models forecast much better, in particular for large samples (T = 3000).

This result is consistent with the established theory which suggests that modelling the volatility as a process which depends only on the past realisations of the returns leads in general to very high orders and is therefore inefficient because it requires many parameters to be estimated.

Nevertheless it must be noted that the estimator of the EGARCH and SV model enjoys an advantage in comparison to the estimator of the EARCH($p$) model: the EKF can deal with nonstationary time series and time varying parameters, while the ordinary LS is based on the assumption of time homogeneity. This assumption is in fact quite restrictive for large samples. Therefore it is probable that an EARCH model

which allows for time varying parameters could show better forecasting performances.

Comparing the performances of the EGARCH and of the SV models leads to the conclusion that these two models are practically equivalent, at least as far as one step forecasting is concerned. Indeed, the values of the criterion sometimes are almost identical. Therefore, for practical application, it can be suggested to consider the predictions of any of these models.

The behaviour of the criterion with respect to the value of $M$ deserves a final remark. First it can be seen that comparing the results for different values of $M$ provides a useful and practical approach to a robust choice of the optimal model. Moreover, we actually find out that different starting points do not influences too much the relative values of the criterion, at least in the data sets that we analysed, so that we can be confident that the assumptions, which underline our model selection strategy represent a good approximation of reality.

| Yen/Dollar exchange rate 01.01.1980-28.06.1991 | | | | | | |
|---|---|---|---|---|---|---|
| Sample size | Beginning of the interval | | | | | |
| | $M = \frac{T}{2}$ | | $M = \frac{T}{3}$ | | $M = \frac{T}{4}$ | |
| $T = 500$ | EARCH(8) | 1276.9 | EARCH(8) | 1604.6 | EARCH(8) | 1816.6 |
| | EGARCH | 1239 | EGARCH | 1567.8 | EGARCH | 1772.9 |
| | SV | 1247.6 | SV | 1570.4 | SV | 1772.9 |
| $T = 1000$ | EARCH(8) | 2233.1 | EARCH(8) | 3122.4 | EARCH(8) | 3510 |
| | EGARCH | 2106 | EGARCH | 2954.8 | EGARCH | 3344.5 |
| | SV | 2103.4 | SV | 2963.9 | SV | 3350.5 |
| $T = 3000$ | EARCH(8) | 6102.7 | EARCH(8) | 8284.9 | EARCH(8) | 9490.2 |
| | EGARCH | 5899.4 | EGARCH | 7936.4 | EGARCH | 8977 |
| | SV | 5876.4 | SV | 7892 | SV | 8932.1 |

Table 4.3: Model selection for the Japanese Yen/US-Dollar exchange rate

| ECU/US-Dollar exchange rate 01.01.1980-28.06.1991 | | | | | | |
|---|---|---|---|---|---|---|
| Sample size | Beginning of the interval | | | | | |
| | $M = \frac{T}{2}$ | | $M = \frac{T}{3}$ | | $M = \frac{T}{4}$ | |
| $T = 500$ | EARCH(9) | 1162.9 | EARCH(9) | 1581.5 | EARCH(9) | 1752.3 |
| | EGARCH | 1131.3 | EGARCH | 1535.7 | EGARCH | 1684.7 |
| | SV | 1124.5 | SV | 1519 | SV | 1662 |
| $T = 1000$ | EARCH(9) | 1978.5 | EARCH(9) | 2698.3 | EARCH(9) | 3141.3 |
| | EGARCH | 1999.7 | EGARCH | 2656.8 | EGARCH | 3124.5 |
| | SV | 1997.6 | SV | 2659.5 | SV | 3116.6 |
| $T = 3000$ | EARCH(9) | 7055.7 | EARCH(9) | 9470.2 | EARCH(9) | 10333 |
| | EGARCH | 7008 | EGARCH | 9310.6 | EGARCH | 10195 |
| | SV | 7010.8 | SV | 9307.5 | SV | 10188 |

Table 4.4: Model selection for the ECU/US-Dollar exchange rate

| Standard and Poor composite 02.01.1986-02.07.1997 | | | | | | |
|---|---|---|---|---|---|---|
| Sample size | Beginning of the interval | | | | | |
| | $M = \frac{T}{2}$ | | $M = \frac{T}{3}$ | | $M = \frac{T}{4}$ | |
| $T = 500$ | EARCH(4) | 1266.5 | EARCH(4) | 1640.8 | EARCH(4) | 1765.8 |
| | EGARCH | 1284.7 | EGARCH | 1626.7 | EGARCH | 1726.2 |
| | SV | 1278.5 | SV | 1630.1 | SV | 1733 |
| $T = 1000$ | EARCH(5) | 2451.5 | EARCH(8) | 3233.8 | EARCH(8) | 3737.1 |
| | EGARCH | 2433.4 | EGARCH | 3238.9 | EGARCH | 3732.2 |
| | SV | 2443.4 | SV | 3238.9 | SV | 3724.6 |
| $T = 3000$ | EARCH(10) | 8149.8 | EARCH(8) | 10514 | EARCH(8) | 11707 |
| | EGARCH | 7964.4 | EGARCH | 10323 | EGARCH | 11537 |
| | SV | 7980.3 | SV | 10365 | SV | 11594 |

Table 4.5: Model selection for the Standard and Poor index

| NIKKEI index 01.01.1980-28.06.1991 | | | | | | |
|---|---|---|---|---|---|---|
| Sample size | Beginning of the interval | | | | | |
| | $M = \frac{T}{2}$ | | $M = \frac{T}{3}$ | | $M = \frac{T}{4}$ | |
| $T = 500$ | EARCH(9) | 1634.2 | EARCH(9) | 1932.1 | EARCH(9) | 2254 |
| | EGARCH | 1465.5 | EGARCH | 1771.3 | EGARCH | 2114.7 |
| | SV | 1447.4 | SV | 1745.6 | SV | 2044 |
| $T = 1000$ | EARCH(9) | 2852 | EARCH(9) | 3953.4 | EARCH(9) | 4486 |
| | EGARCH | 2435.1 | EGARCH | 3389.8 | EGARCH | 3390.2 |
| | SV | 2452.9 | SV | 3384.3 | SV | 3890.4 |
| $T = 3000$ | EARCH(10) | 9208.6 | EARCH(10) | 11862 | EARCH(10) | 13207 |
| | EGARCH | 7661.3 | EGARCH | 9962.4 | EGARCH | 11106 |
| | SV | 7528.2 | SV | 9837.5 | SV | 10207 |

Table 4.6: Model selection for the NIKKEI index

| Volkswagen stock price 01.01.1980-28.06.1991 | | | | | | |
|---|---|---|---|---|---|---|
| Sample size | Beginning of the interval | | | | | |
| | $M = \frac{T}{2}$ | | $M = \frac{T}{3}$ | | $M = \frac{T}{4}$ | |
| $T = 500$ | EARCH(4) | 736.59 | EARCH(4) | 1042.6 | EARCH(4) | 1255.9 |
| | EGARCH | 633.43 | EGARCH | 930.27 | EGARCH | 1100.7 |
| | SV | 634.76 | SV | 930.99 | SV | 1055.2 |
| $T = 1000$ | EARCH(4) | 1823.7 | EARCH(4) | 2296.3 | EARCH(4) | 2560.2 |
| | EGARCH | 1693.4 | EGARCH | 2081.7 | EGARCH | 2314.6 |
| | SV | 1690.9 | SV | 2077.9 | SV | 2312.1 |
| $T = 3000$ | EARCH(9) | 6647.3 | EARCH(8) | 8520.4 | EARCH(8) | 9490.2 |
| | EGARCH | 6167.7 | EGARCH | 7901.9 | EGARCH | 8738.1 |
| | SV | 6116.8 | SV | 7845.2 | SV | 8679.4 |

Table 4.7: Model selection for the Volkswagen stock prices

### 4.3.3 A graphical example: the NIKKEI stock index

Here the plots of the estimation for the EGARCH and SV models for the NIKKEI data set are presented. The analysis of the graphical output underlines the similarities in the behaviour of these models. The plots of the forecasted volatility (Figure 4.4) and the ones of the autocorrelation of the forecast error (Figure 4.7) are almost identical for both models. The observation of the ACF is also very interesting because we see that the residuals have not been fully whitened by the estimators, so that EGARCH and SV models of higher order could eventually improve the forecasting performance.

The plots of the recursive estimates of the parameters are also presented in Figures 4.6 and 4.5. One must be careful when interpreting those figures, because the extended Kalman filter may give a excellent estimate of the volatility $h_t$, even when the parameters are time varying. However the estimate of the parameters may not be so good (Chui & Chen 1998, pagg. 124-128). Indeed, one can see a shift in the volatility level from the plot of the returns (Figure 4.1), where more or less after the 2500th the average deviations from the mean becomes larger. This assumption is confirmed by the plots of the volatility (Figure 4.4) and also by the estimates of the intercept, which recognises very quickly the change point. Actually not only the intercept has an adjustment, but also the other parameters. Nevertheless it is not clear if this happens because an adjustment is needed, or because of the correlation between the estimators. We can conclude that, as far as the parameters are concerned, the extended Kalman filter provides a very flexible estimation technique. On the other hand, this flexibility leads to a certain variability of the estimator.

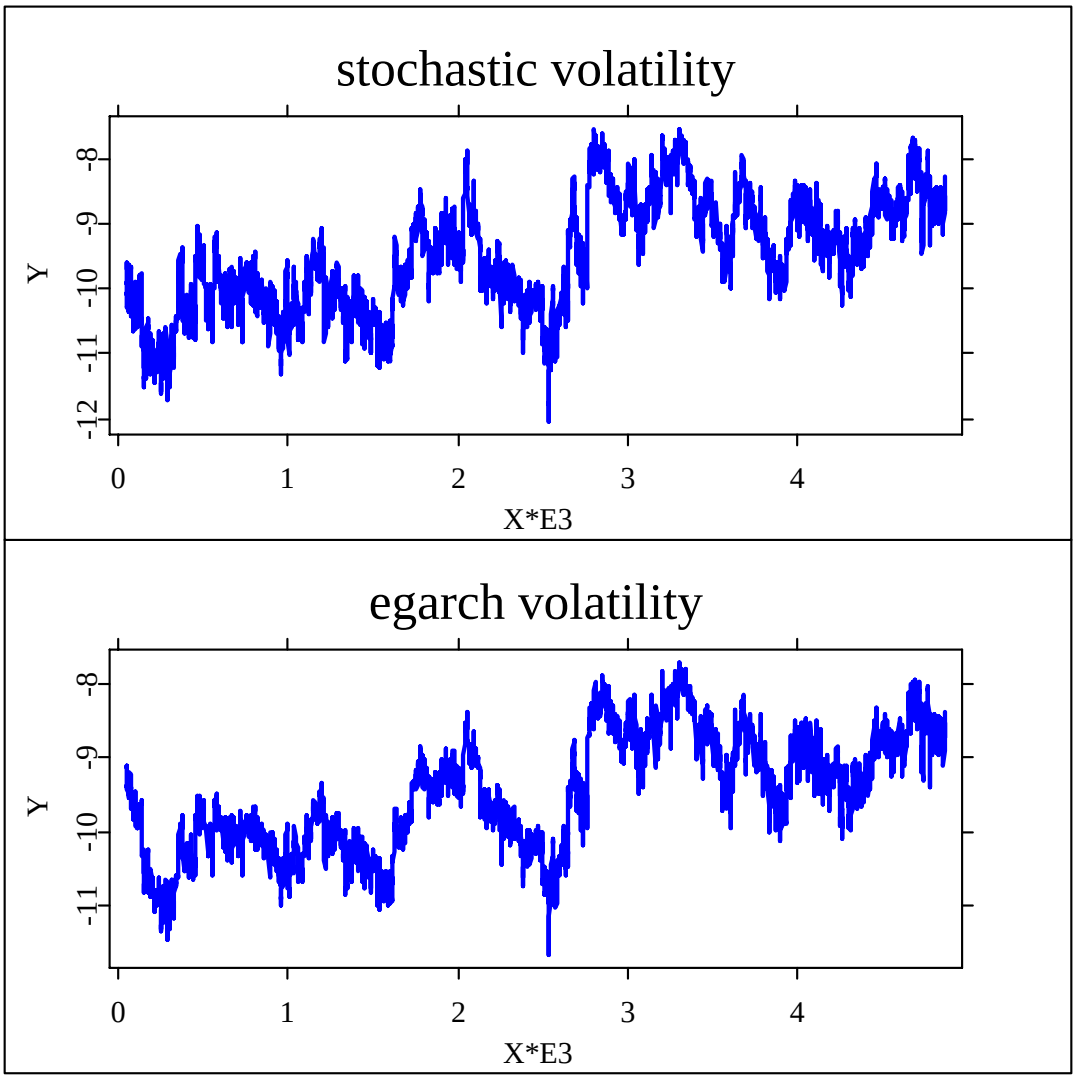


Figure 4.4: Forecasted log-volatility of the returns of the NIKKEI stock index

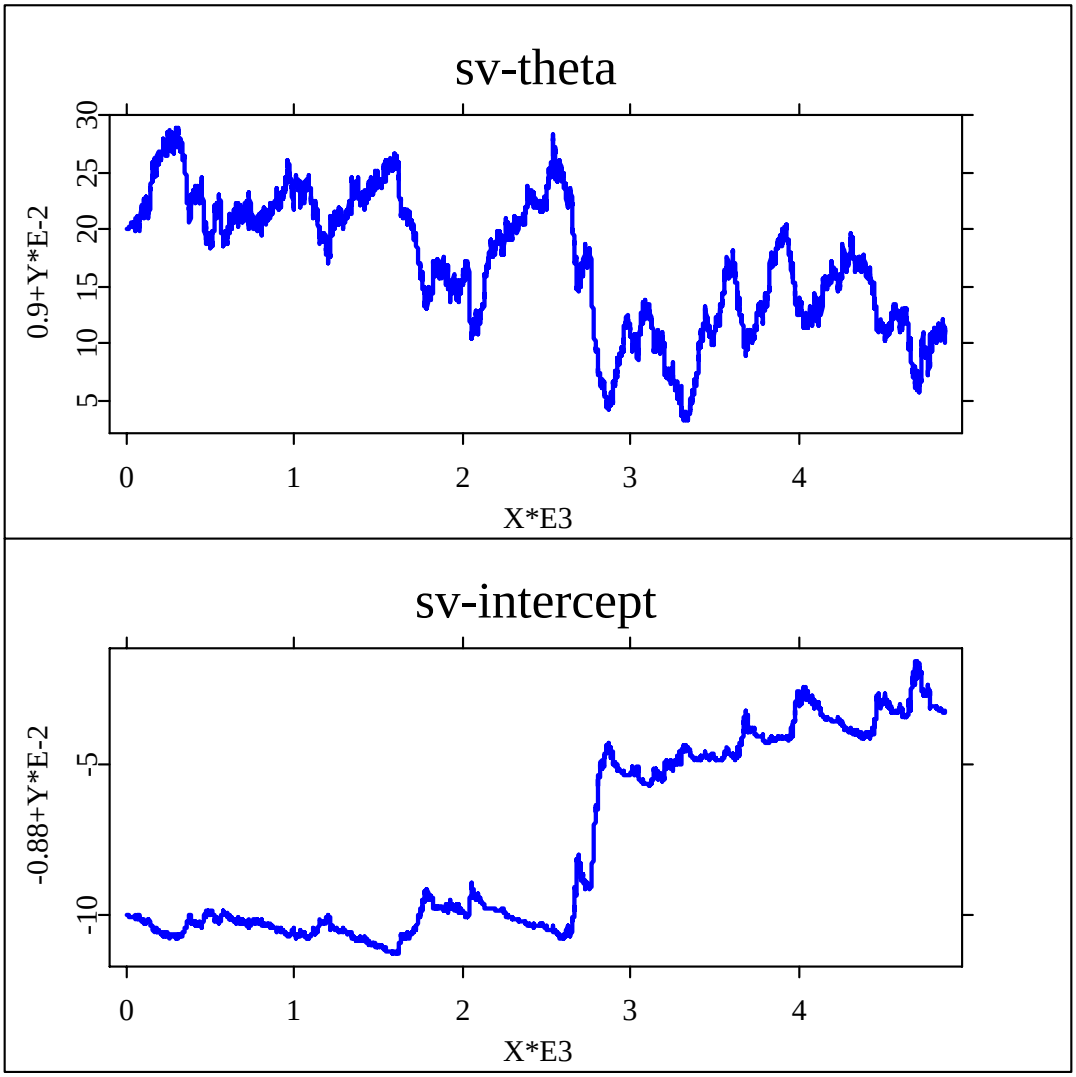


Figure 4.5: Estimations of the parameters of the stochastic volatility model for the NIKKEI stock index

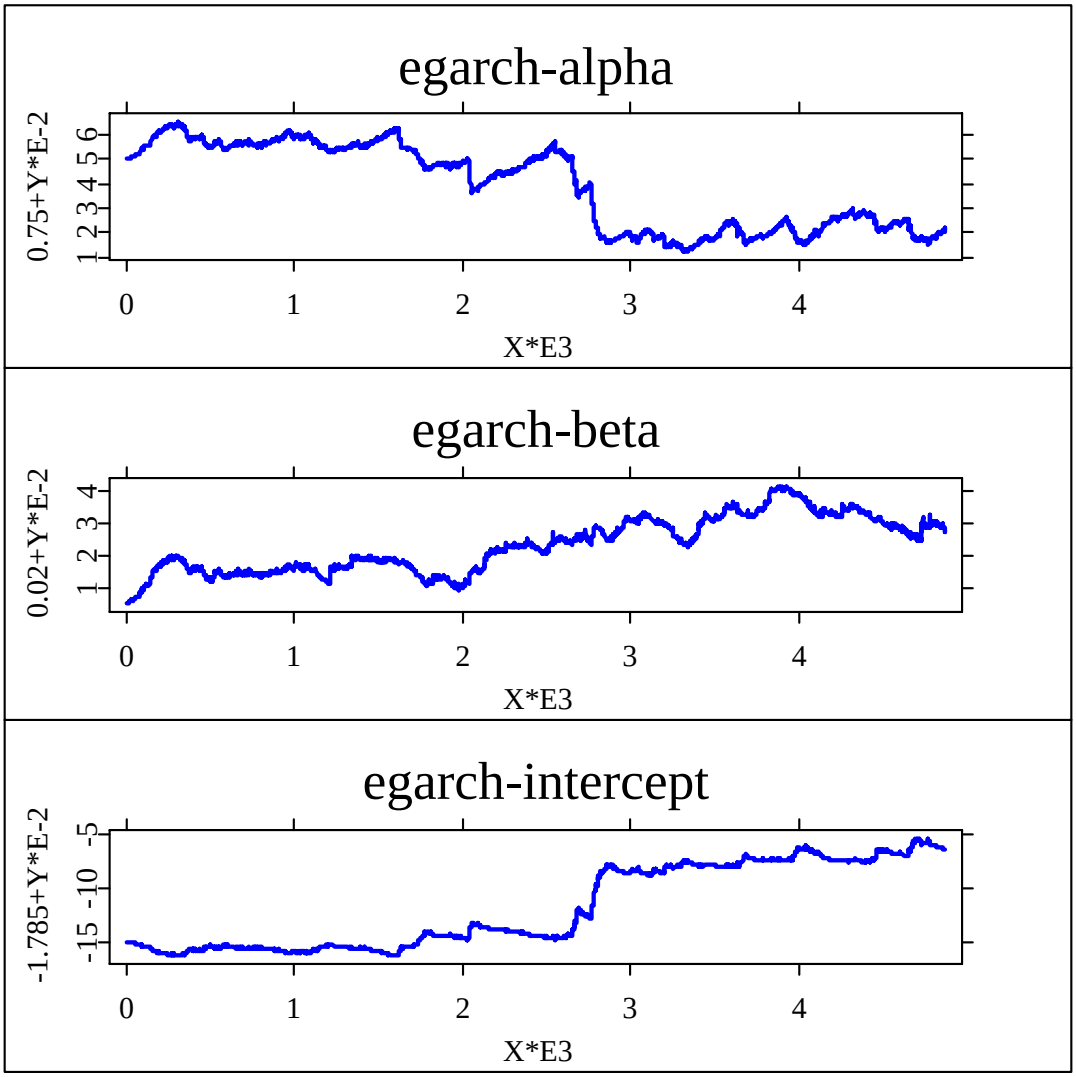


Figure 4.6: Estimations of the parameters of the EGARCH model for the NIKKEI stock index

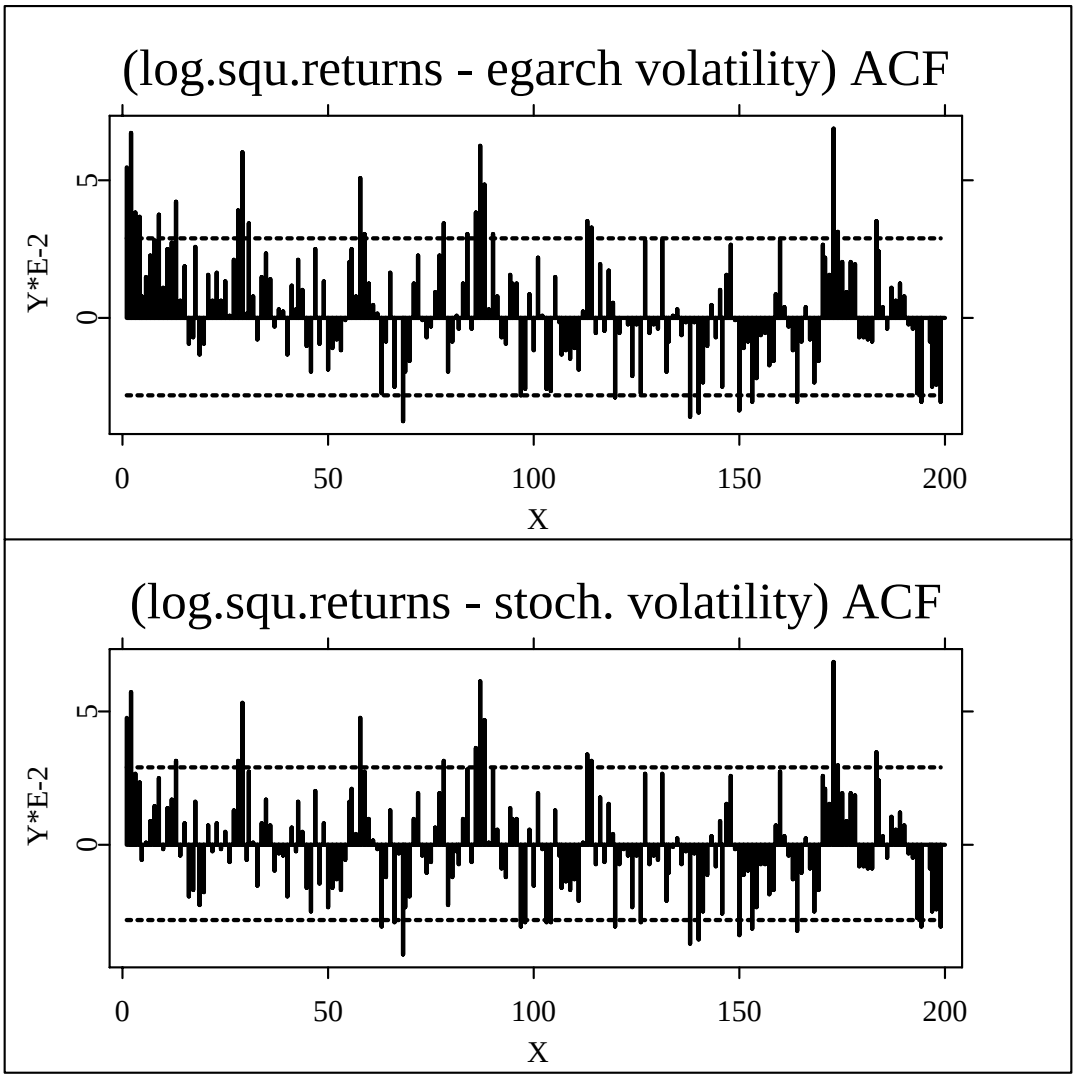


Figure 4.7: Autocorrelations of the log-squared returns minus the forecasted log-volatility for the NIKKEI stock index

### 4.3.4 Forecast confidence bands

One possible application of the volatility models consists in building confidence forecast bands. Recall the three basic equations which describe our model for a financial asset: (4.1), (4.2) and (4.3). Under these assumptions, the best forecast of the price (or for the log-price) at time $t$, is the price (the log price) at time $t-1$. Let us consider the log-prices, in this case the forecast error is:

$$\ln P_t - \ln P_{t-1} = \sigma_t \varepsilon_t,$$

the expected forecast error is zero, and the forecast error variance is $\sigma_t^2$. Keeping the assumption that $\varepsilon_t$ has a standard normal distribution, one can construct confidence intervals for $\ln P_t$, for example a 95% forecast interval:

$$\left[\ln P_{t-1} - 1.96\sigma_t,\ \ln P_{t-1} + 1.96\sigma_t\right],$$

or equivalently for the prices:

$$\left[P_{t-1}/\exp(1.96\sigma_t),\ P_{t-1}\exp(1.96\sigma_t)\right].$$

The actual value of $\sigma_t$ is unknown, but in the previous sections we have analysed some methods for the forecasting of the volatility. In particular, we have focused on the log-volatility $h_t$, and we have considered some forecasting methods, so that at each point time we are able to produce a forecast for the next period. If $\hat{h}_{t|t-1}$ is the forecast of the log-volatility for time $t$, the forecast for the standard deviation is:

$$\hat{\sigma}_{t|t-1} = \sqrt{\exp\left(\hat{h}_{t|t-1} + 1.27\right)},$$

and the estimated confidence interval becomes:

$$\left[P_{t-1}/\exp(1.96\hat{\sigma}_{t|t-1}),\ P_{t-1}\exp(1.96\hat{\sigma}_{t|t-1})\right].$$

For each data set we compute the value of the forecast bands for the SV and EGARCH model. Table 4.8 shows the ratio of the number of points which lay outside the confidence bands, divided by the total number of observations. All these ratio are very close to 0.05, which represents the theoretical value which is reached under optimal conditions, so that the results are very satisfactory.

It is interesting to see that the empirical frequency with which the observations lay outside the confidence bands is always a bit larger than the theoretical one. More than 5% of the observations lay outside the confidence bands, which means that they are in general a little too narrow, so that one can presume a small downward bias in the estimates of the volatility. This bias may be due to the fact that the distribution of $\ln \varepsilon_t^2 = v_t$ is not symmetric and presents a negative skewness, not only if $\varepsilon_t$ is normally distributed, but for very general distribution assumptions.

The confidence bands for the NIKKEI stock index are displayed in Figure 4.8. The data have been zoomed to make observation possible. One can clearly recognise how the confidence bands is wider or narrower to adapt to the periods of high and low variance.

| Empirical alpha for the confidence bands | | |
|---|---|---|
| NIKKEI | EGARCH | 0.08073 |
| | SV | 0.07765 |
| ECU/US-Dollar | EGARCH | 0.06840 |
| | SV | 0.06820 |
| Yen/US-Dollar | EGARCH | 0.07452 |
| | SV | 0.07361 |
| Volkswagen | EGARCH | 0.05710 |
| | SV | 0.05731 |
| Standard and Poor | EGARCH | 0.08011 |
| | SV | 0.08463 |

Table 4.8: Frequence of the observations which do not lie within the confidence bands relative to the whole sample

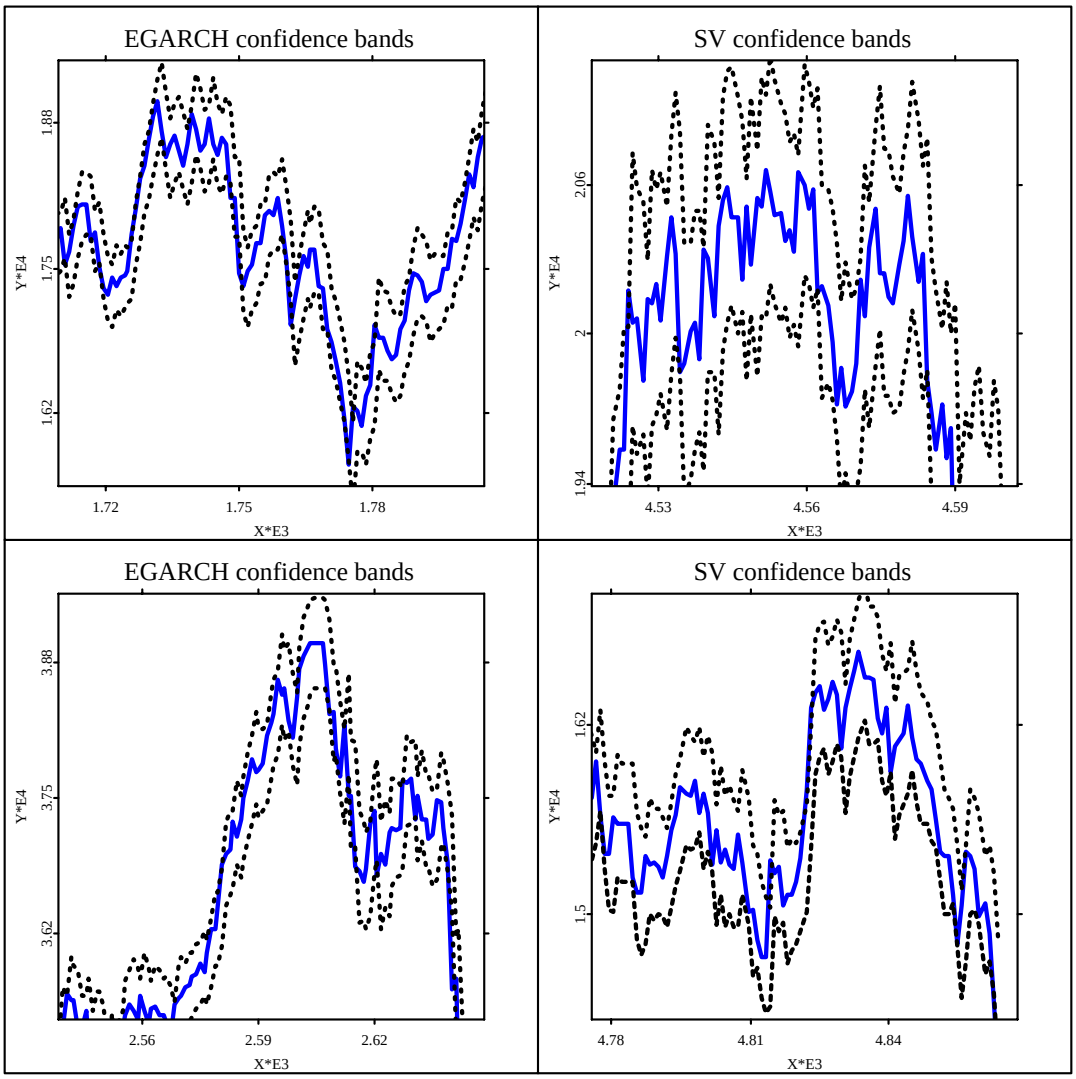


Figure 4.8: Zoom of the forecast confidence bands for the NIKKEI stock index

# Conclusion

In this study a new approach to model selection was proposed. The basic idea underlying this approach is that the true data generating process is unknown and in practice cannot be recovered. Therefore, one needs an approximation of the true process which satisfies some optimality criterion. In particular, the focus is set on one-step-forecasting and on the minimisation of the forecast uncertainty, which is a relevant concern for practical applications. The optimal model is defined as the one which minimises the mean square forecast error, while the estimator of the optimal model is the one that minimises the sum of the square forecast errors.

This model selection strategy appears to be easy to implement, and allows for the comparison of very different models without requiring strong assumptions such as stationarity. Simulation results for autoregressive processes and a practical application on financial data are very satisfactory. In particular the choice of $M$ (the starting point for the summation of the square forecasting residuals) appears to have no strong influence on the results, and the simple comparison of the values of the criterion for some values of $M$, is sufficient in casting eventual doubts about which model to select. Unfortunately only very few results about the theoretical properties of the model selection criterion could be presented and this topic deserves further investigation.

In a practical application three models for the volatility of financial time series were compared: EARCH($p$), EGARCH and SV. The EARCH($p$) model is outperformed by the other two in particular for large samples. Nevertheless it must be noted that the estimation technique for the EGARCH and SV model accounts for parameter changes, while the one for the EARCH($p$) model does not. Therefore it would be interesting to make a further comparison with an EARCH($p$) model which allows for time varying parameters. The performances of the EGARCH and SV models are very similar, so that any of these models may be recommended. Nevertheless some interesting questions remains open, concerning the behaviour of other estimators, larger orders for the SV and EGARCH models, and in particular the applicability of these techniques for a real problem of the financial praxis, such as option pricing or hedging.

# Bibliography


Aoki, M. (1990). *State Space Modeling of Time Series*, Springer-Verlag, Berlin.

Basci, S. and Zaman, A. (1998). Variance estimates and model selection, *Technical report*, Bilkent University, Ankara.

Bollerslev, T. (1995). Generalised autoregressive conditional heteroskedasticity, *in* Engle (1995b).

Bollerslev, T. P., Chou, R. Y. and Kroner, K. F. (1992). Arch modeling in finance: A review of the theory and empirical evidence, *Journal of Econometrics* **31**: 309–328.

Chui, C. and Chen, G. (1998). *Kalman Filtering*, Information Sciences, third edn, Springer-Verlag, Berlin.

Danielsson, J. (1994). Stochastic volatility in asset prices estimation with simulated maximum likelihood, *Journal of Econometrics* **64**: 375–400.

Elliot, R. J., Aggoun, L. and Moore, J. B. (1995). *Hidden Markow Models*, Springer-Verlag, Berlin.

Engle, R. F. (1995a). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation, in *ARCH, selected readings* (Engle, 1995b).

Engle, R. F. (ed.) (1995b). *ARCH, selected readings*, Oxford University Press, Oxford.

Feldmann, D. (1998). *The costs of delta-hedging for heteroskedastic volatility models*, Master's thesis, Humboldt University Berlin.

Franses, P. and Dijk, D. V. (1996). Forecasting stock market volatility using (non-linear) garch models, *Journal of Forecasting* **15**: 229–235.

Gouriéroux, C. (1997). *ARCH Models and Financial Application*, Springer-Verlag, Berlin.

Hafner, C. and Herwartz, H. (1997). Structural analysis of portfolio risk using beta impulse response functions, *Technical report*, Sonderforschungsbereich 373, Humboldt University Berlin.

Hafner, C. M. (1998). *Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility*, Physica-Verlag, Heidelberg.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press, Princeton.

Härdle, W. and Hafner, C. M. (1997). Discrete time option pricing with flexible volatility estimation, *Technical report*, Sonderforschungsbereich 373, Humboldt University, Berlin.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, Cambridge.

Harvey, A. C. (1992). *Time Series Models*, second edn, Harvester Wheatsheaf, New York.

Harvey, A., Ruiz, E. and Shephard, N. (1995). Multivariate stochastic variance models, *in* Engle (1995b).

Hendry, D. F. (1995). *Dynamic Econometrics*, Oxford University Press, Oxford.

Hull, J. and White, A. (1987). The pricing of option on assets with stochastic volatilities, *The Journal of finance* **42**: 281–302.

Jacod, J. and Shiryaev, A. N. (1987). *Limit theorems for stochastic processes*, number 288 in *Grundlehren der Mathematischen Wissenschaften*, Springer-Verlag, Berlin.

Jacquier, E., Polson, N. G. and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models, *Journal of Business and Economic Statistics* **12**: 371–417.

Johnson, H. and Shanno, D. (1987). Option pricing when the variance is changing, *Journal of Financial and Quantitative Analysis* **22**: 419–437.

Lewis, R. and Reinsel, G. (1982). Prediction of multivariate time series by autoregressive model fitting, *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, pp. 144–149.

Lütkepohl, H. (1986). *Forecasting Aggregated Vector ARMA Processes*, Springer-Verlag, Berlin.

Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*, second edn, Springer-Verlag, Berlin.

Monfardini, C. (1996). Estimating stochastic volatility models through indirect inference, *Technical report*, European Universitary Institute Florence.

Nelson, D. B. (1995). Conditional heteroscedasticity in asset returns: A new approach, *in* Engle (1995b).

Scott, L. O. (1987). Option pricing when the variance changes randomly: Theory, estimation, and an application, *Journal of Financial and Quantitative Analysis* **22**: 143–151.

Singer, H. (1998). *Finanzmarktökonometrie: Zeitstetige Systeme und ihre Anwendung in Ökonometrie und empirischer Kapitalmarktforschung*, Physica-Verlag, Heidelberg.