Econometric Forecasting

Robert M. Kunst

September 2012

1 Introduction

This lecture is confined to forecasting in the context of time and, consequently, time series (*time-series forecasting*). There are also forecasting problems outside the time context, for example if k out of n data points are observed and the remaining $n - k$ points are to be 'predicted'. In a wide sense, all extrapolations can be viewed as forecasts. However, we will consider only forecast problems with the following structure:

1. one or several variables are observed over a time period $t = 1, \ldots, N$: $x_1, \ldots, x_N$;
2. a 'future' data point x_{N+h} is to be approximated using the observations.

For this problem, the English language uses the words *forecast* (Teutonic origin, to cast or throw ahead) and *prediction* (Latin origin, to tell in advance, cf. German '*vorhersagen*'). Some authors try to separate the two words by definition, usually using *forecast* for the aspect of substance matter in a theory context, where the aim is a complete vision of a future state, and using *prediction* for the purely mathematical and statistical aspect. The forecasting literature tends to use the participle '*forecasted*', which is grammatically incorrect but convenient. The word *prognosis* (Greek origin, to know in advance) is used only in the context of medicine, while '*Prognose*' is commonly used in the German language. Sometimes, the word *projection* (Latin origin, literally equivalent to 'forecast') appears as a synonym for 'forecast', while it unfortunately has a specific meaning in mathematics that is not necessarily related to prediction.

Possible aims of forecasting are:

1. Interest in the future

2. Planning
3. Control
4. Risk management
5. Evaluation of different policies
6. Evaluation of different models

An example for planning is the management of inventories on the basis of forecasts for the demand for certain goods. Control and policy evaluation are complex tasks, as they admit a feedback from future action, which may react to the forecasts, to the predicted variables. Economists may use words such as *conditional forecast*, *policy simulation*, or *scenarios*. Currently, the aim of controlling an economy is often rejected explicitly, as there is a widespread aversion to 'controlled economies'. The aim of policy evaluation ('simulation') is sensitive to the argument of the Lucas critique (the reaction mechanism of the economic system is affected by changing seemingly exogenous variables). Also model evaluation is a problematic aim, as good models with high explanatory power regarding actual cause-effect relationships are not necessarily the best forecast models. Some forecasters renounce forecasting out of pure interest in the future ('forecasts without any decision based on them are useless'), although it reflects an innate human impulse (curiosity). This impulse has left its marks throughout the history of mankind. It is also typical of individual human evolution. In the first decade of their lives, human beings are not yet fully conscious of the differences among the past (known, but subject to forgetting), the present (immediate sensory perception) and the future (unknown). Children often assume that lost objects, demolished buildings, deceased persons will re-appear in the future, although 'more rarely'. Only the full realization of the passing of time incites the search for methods to know the future 'in advance'. Early cultures may have focused on magical procedures of forecasting. It is interesting to consider the quotation by Wayne Fuller:

> *The analysis of time series is one of the oldest activities of scientific man.*

Possibly, Fuller was thinking of prehistoric priests who investigated the movement of celestial bodies, in order to advise hunters or farmers. That type of time-series analysis almost certainly targeted prediction.

Excluding magical procedures ('magical' = no recognizable cause-effect relationship between data and event to be forecast), Chris Chatfield divides forecasting procedures into three groups:

1. subjective procedures (*judgemental forecasts*), such as questioning experts and averaging the collected response (so-called Delphi method), which are also applied to economic short-run forecasts. However, one cannot rule out that the questioned experts are using models.

2. *univariate* procedures use data on a single variable, in order to predict its future

3. *multivariate* procedures use several variables jointly (a vector of variables)

Furthermore, one could distinguish: *model-free* procedures, which are often subsumed under the heading of 'extrapolative procedures' and do not attempt to describe explicitly the dynamics of the variables; and *model-based* procedures, which entertain models of the 'data-generating process (DGP)', often without assuming that these models explain the actual cause-effect mechanisms. Both groups comprise univariate as well as multivariate methods. Typical model-free procedures are *exponential smoothing* and *neural nets*. Typical model-based procedures are time-series analytic procedures and also macroeconometric modelling.

Chatfield further distinguishes *automatic* and *non-automatic* methods. An automatic method generates forecasts on the basis of a well-defined rule without any further intervention by the user. Some forecasting procedures in business economics and finance must be automatic, due to the large number of predicted variables. Most published economic forecasts are non-automatic. Even when models are utilized for their compilation, experts screen the model predictions and modify them so that they conform to their subjective view (Ray Fair talks about 'add factors', if subjective constants are added to automatic forecast values, others are using the word 'intercept correction').

self-defeating and self-fulfilling forecasts: often, decision makers react to forecasts. This reaction, which is usually not anticipated in the model that is used in generating the forecast, may imply that forecasts become correct, even when they would otherwise be of low quality. For example, a forecast of inflation may lead to actual inflation, by way of the wage-price spiral and wage bargaining (*self-fulfilling*). Conversely, a predicted high unemployment rate may entail political action to curb unemployment, such that the predicted high unemployment rate will not materialize (*self-defeating*). Depending on the aim of the forecast, it may make sense to anticipate the reaction of decision makers in forecasting (endogenous policy). In a longer-run economic scenario, this kind of anticipation is a basic requirement.

Many text books report curious examples for incorrect forecasts. For example, the (correct) prediction of increasing traffic in London has caused the

prognosis of gigantic quantities of horse manure in the streets, some decades later the prognosis of endless networks of rails and suffocating coal steam. TV sets and PC's rank among the classical examples for technological inventions, whose impact on the market was underestimated severely, while image telephones and flying machines were overestimated. Such curiosities can be explained *ex post* at least partly by forecasters' inability to account for yet unknown technological developments (automobiles, electrical vehicles for public transportation), partly by the always extremely difficult assessment of consumers' acceptance of innovations. It should be added that forecasts for the volumes and directions of international trade have proven surprisingly accurate, and this is also true for demographic projections. A good example for a remarkable improvement in forecasting accuracy are weather forecasts, whose uncertainty has diminished significantly in the last decades. Unfortunately, no similar remark can be made regarding short-run macroeconomic forecasts.

out-of-sample and in-sample forecasts: A forecast, in symbols

$$\hat{x}_N(h),$$

i.e. a forecast that is compiled at time N to approximate the yet unknown value x_{N+h}, should use only information that is available at time N. Then, the forecast is called *out-of-sample*. The available information consists of the *data* $x_N, x_{N-1}, \ldots$ and possibly a forecast model, which is completely determined from the available information. Free parameters of the forecast model should be estimated from the past only, and even model selection should rely on past information. Even in academic publications, many reportedly 'out-of-sample' forecasts do not fulfil these conditions. Conversely, the *in-sample* forecast uses a forecast model that relies on the entire sample and predicts observations within the sample range. The prediction errors of such *in-sample* forecasts are simply residuals that yield no information on the predictive accuracy of the procedure.

Particularly the German literature uses the expressions *ex-ante* and *ex-post*. Their definition varies across authors. They can be equivalent to *out-of-sample* and *in-sample*, or *ex-post* may be *out-of-sample* in some of its aspects, and *in-sample* in others. For example, one may be interested in forecasts that use data on 'endogenous' variables and also estimates from the time period preceding N, while 'exogenous' variables are assumed as known for the time range $N+1, N+2, \ldots, N+h$. In macroeconomics, one often distinguishes the original data points that were known physically at time N and later data revisions that are also indexed by N. Such intermediate forms are of limited validity with respect to an assessment of predictive accuracy.

2 Model-free methods of extrapolation

The simplest prediction methods *extrapolate* a given sample beyond the sample end N. For example, this could be done by hand ('free-hand' method) in a time-series plot. A graph with time on the x–axis and the variable on the y–axis is an important visual tool. CHATFIELD calls this representation the *time plot*. Usually, computer programs admit the options either to represent the data as unconnected points (symbols) or as a connected broken line or as a broken line with added symbols. Observations are available at *discrete time points* only, therefore the unconnected representation appears to be more 'honest'. However, connected representations tend to be more pleasing to the human eye. They may also ease the recognition of important patterns.

Whereas generally data are available in discrete time only—usually with equidistant intervals between time points—we may distinguish several cases that are also relevant in economics (following CHATFIELD):

1. the variable exists *in continuous time*, while it is only measured and reported at certain time points. Examples are measurements of temperature or certain prices of frequently traded assets on stock exchanges, including exchange rates, or the population. In principle, it is conceivable to increase sampling frequency for these cases.

2. the discretely measured variable exists only at certain time points, for example wages and salaries.

3. the discretely measured variable is a flow variable that is defined from time aggregation only, for example gross domestic product (GDP) or most macroeconomic variables. The GDP measures transactions within a time interval and not at a time point.

For the last case, connecting data points seems incorrect, although it may help with visual interpretation. The distinction of the cases is important for the problem of *time aggregation*, if the forecaster decides to switch to a lower observation frequency, for example from quarters to years. For this genuinely economic problem, there are again three cases:

1. measurements at well-defined time points, for example a monetary aggregate at the end of a month. In this case, it is easy to switch to a lower observation frequency, for example to the money aggregate at the end of a quarter.

2. averages over a time interval. For example, a quarterly average is defined from the average over three monthly averages for non-overlapping months.

3. Flows, whose quarterly values should be the sum of three monthly values.

A popular alternative to free-hand forecasting are the many variants of *exponential smoothing*. While exponential smoothing can be justified by being optimal under certain model assumptions, the method is not really based on the validity of a model. Literally, smoothing is not forecasting. Systems analysis distinguishes among several methods that are used for recovering 'true underlying' variables from noisy observations. *Filters* are distinguished from *smoothers* by the property that filters use past observations only to recover data, while smoothers use a whole time range to recover observations within the very same time range.

2.1 Exponential smoothing

The simplest smoothing algorithm is *single exponential smoothing* (SES). In SES, the filtered value at t is obtained by calculating a weighted average of the observations and the filtered value at $t-1$

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

where $\alpha \in (0,1)$ is a *damping* or *smoothing* factor. This equation is also called the 'recurrence' form. Small α implies strong smoothing. Chatfield also mentions the equivalent 'error-correction' form (see also Makridakis *et al.*)

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

where $x_t - \hat{x}_{t-1}$ can be interpreted as the prediction error at t. Alternatively, by repeated substitution one obtains

$$\hat{x}_t = \alpha \sum_{j=0}^{t-1} (1-\alpha)^j\, x_{t-j},$$

which apparently assumes $\hat{x}_0 = 0$ as a starting value.

All exponential smoothing algorithms face two problems in application. The first is how to determine parameters, such as the smoothing factor α. The other one is how to determine starting values for unobserved variables, such as $\hat{x}_0$. For the parameters, there are two options:

1. Fix the constants *a priori*. The literature recommends to pick α from the range $(0.1, 0.3)$. It may be inconvenient to choose as many as three parameters arbitrarily, which are required by some of the more complex variants.

2. Estimate parameters by fitting a model to data. This approach is recommended by many authors (e.g. GARDNER 'The parameters should be estimated from the data'), although it is at odds with the idea of a model-free procedure. Estimation—or 'minimization of squared errors' $\Sigma\left(x_{t+1} - \hat{x}_t\right)^2$, as it is described in computer manuals—requires a computer routine for nonlinear optimization, which may be cumbersome, unless the specific exponential-smoothing variant is known to be equivalent to model-based filtering. In that case, a simple time-series model can be estimated to yield the parameters.

Similar problems concern the starting values. Here, three solutions are offered:

1. Fix starting values by fitting simple functions of time to data, in the spirit of the respective algorithm. For example, SES should be started from the sample mean as a starting value, while variants that assume a trend would fit a trend line to the sample, for example by simple least-squares fitting.

2. Start from the observed data, such as $\hat{x}_1 = x_1$. If α is small and the sample is not too large, this choice may result in severe distortions.

3. Use *backcasting*. The recursion equation is used to determine $\hat{x}_1$. Instead of running it from 1 to N, one runs the recursion from N to 1, using some starting value for $\hat{x}_N$.

Historically, the method was introduced as a solution to estimating a level or expectation of x_t, if that level changes slowly, although without specifying a formal model for that changing level. A suggested solution was to replace least squares (the mean or average) by discounted least squares, i.e. to minimize

$$\sum_{j=0}^{t-1} \beta^j \left(x_{t-j} - \mu\right)^2,$$

which yields least squares for $\beta = 1$. For $\beta = 1 - \alpha$, SES is obtained.

Example. Figure 1 shows the application of SES to a time series of Austrian private consumption. This version of exponential smoothing uses the average over the first half of the sample as starting value $\hat{x}_1$.

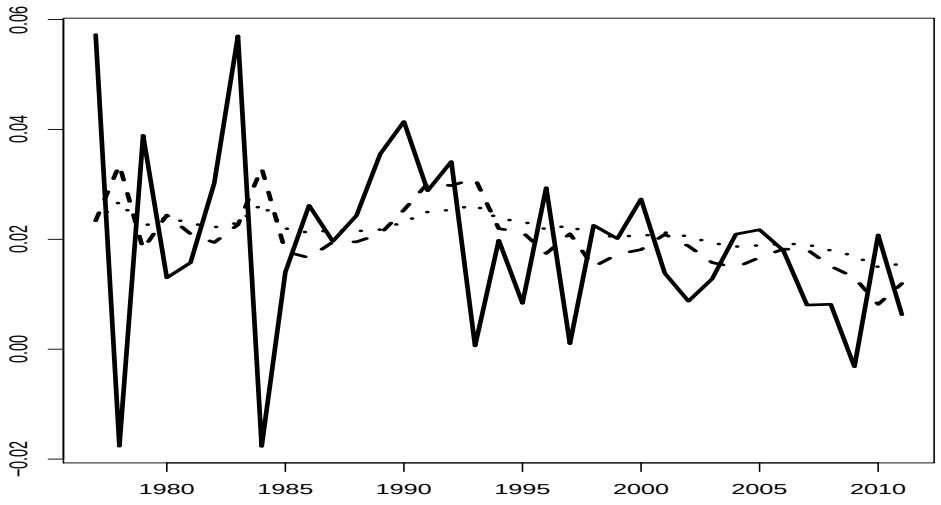

Figure 1: Growth rates of Austrian private consumption 1977–2011 at constant prices and *single exponential smoothing*. Solid curve is the data, short dashes are SES values at $\alpha = 0.1$, long dashes are SES values at $\alpha = 0.3$.

The *forecast* from exponential smoothing is defined by simply extrapolating $\hat{x}_N$ flatly into the future, thus, in CHATFIELD's notation.

$$\hat{x}_N(1) = \hat{x}_N(2) = \ldots = \hat{x}_N.$$

Clearly, the procedure is inappropriate for trending data, which are the rule in economics. For trending data, a first suggestion is to run the SES method twice. This is called *double exponential smoothing* (DES) with the recursions

$$\begin{aligned} L_t &= \alpha x_t + (1-\alpha) L_{t-1}, \\ T_t &= \alpha L_t + (1-\alpha) T_{t-1}. \end{aligned}$$

L_t is meant to denote a local level estimate, while T_t is a local trend estimate. However, note that T as well as L follow the trend in the data when such a trend exists. This method apparently assumes that the data lie on a trend line, with its intercept and slope changing slowly. Forecasts are obtained from extrapolating the last observations on L_t and T_t:

$$\begin{aligned} \hat{x}_N(h) &= \left(2 + \frac{\alpha h}{1-\alpha}\right) L_N - \left(1 + \frac{\alpha h}{1-\alpha}\right) T_N \\ &= 2L_N - T_N + \frac{\alpha}{1-\alpha}(L_N - T_N)\, h \end{aligned}$$

CHATFIELD maintains that the method is not so often applied nowadays. Nevertheless, it is still contained in many computer programs. Like SES, it only requires one smoothing parameter. Convenient starting values L_1 and T_1 can be obtained from regressing the sample on a linear function of time. $\hat{x}_{t-1}(1)$ may be seen as a 'smoothed' version of x_t. Within the sample, it is not a true forecast, as it uses information from the whole sample regarding starting values and possibly the parameter α.

It can be shown (see Section 3.2) that SES corresponds to a model-based forecast for the ARIMA$(0,1,1)$ model given by

$$\Delta X_t = \varepsilon_t + (\alpha - 1)\,\varepsilon_{t-1}.$$

These models are generalizations of the random walk $\Delta X_t = \varepsilon_t$. They are usually considered for the range $\alpha \in (0,2)$. For $\alpha = 1$, we obtain a random walk and the optimal forecast is the last value. For $\alpha > 1$, the ARIMA model would yield a forecast that cannot be obtained by SES, as that method restricts α to $(0,1)$. A different model class, for which SES is optimal, are models of contaminated random walks that are defined from random walks with added noise. Because these contaminated models are again members of the above ARIMA class, they do not need to be considered separately.

DES can be shown to be optimal for certain ARIMA$(0,2,2)$ models of the form

$$\Delta^2 X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2},$$

where θ_1 and θ_2 are subject to the restriction that $1+\theta_1 z + \theta_2 z^2 = (1+\zeta z)^2$ for some ζ. This is a relatively implausible restriction, and the modelling of second-order differences may also be implausible.

Example. Figure 2 shows the application of DES to the Austrian consumption data whose growth rates were used in Figure 1. Because the original data 'in levels' are clearly trending, DES may be appropriate. The program EViews determines the starting value in year 1 by fitting a linear trend regression to the first half of the sample. 'Estimation' of α would imply a value of 0.36 in this example, intermediate between the close fit of $\alpha = 0.5$ and the loose fit of $\alpha = 0.1$ that are shown in the graph.

2.2 Holt-Winters methods

Holt-Winters methods are generalizations of exponential smoothing. They cover cases of trending as well as seasonal data. Typically, Holt-Winters algorithms need more than one smoothing parameter. The simplest one is *Holt's linear trend method*, otherwise known as Holt-Winters no-seasonal

Figure 2: Austrian private consumption 1976–2011 and *double exponential smoothing* (DES). Solid curve marks data, short dashes DES values at $\alpha = 0.1$ and long dashes DES values at $\alpha = 0.5$.

method. Holt's method defines a *local trend* T_t as well as a *local level* L_t by the following recursions

$$\begin{aligned} L_t &= \alpha x_t + (1-\alpha)\left(L_{t-1} + T_{t-1}\right), \\ T_t &= \gamma\left(L_t - L_{t-1}\right) + (1-\gamma)\, T_{t-1}. \end{aligned}$$

The local level is a weighted average of the observation and the value 'predicted' from the last level and trend estimates. The local trend or slope is a weighted average of the increase in the local level and the last slope. Convenient starting values may be $L_1 = x_1$ and T_1 determined as the sample mean of the first differences of the data. Note that the meaning of T_t differs from the DES method.

Holt's procedure defines forecasts by

$$\hat{x}_N(h) = L_N + hT_N,$$

such that $\hat{x}_{t-1}(1) = L_{t-1} + T_{t-1}$ can be viewed as a smoothed version of x_t. While the procedure is motivated by being appropriate for data with slowly changing local linear trends, it can be shown that it is optimal for certain

ARIMA$(0, 2, 2)$ models, that is, for models such as

$$\Delta^2 X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}.$$

If a time series follows this model, it becomes only stationary after taking second-order differences. In economics, this second-order integratedness was only found for some price series and monetary aggregates. Nevertheless, if merely seen as a forecasting procedure, Holt's linear trend method appears to give good forecasts for many economic series.

Example. Figure 3 shows the application of the Holt method to the Austrian consumption data. For larger α and small γ, we obtain a reasonable fit, while a loosely fitting curve trend is implied by large γ. The curves were obtained by EViews, which uses true data values as starting values for the first time point here. This explains the deteriorating fit, as time progresses.

Figure 3: Austrian private consumption and *Holt's linear trend method*. Solid curve shows data, short dashes mark Holt values for $\alpha = 0.5$ and $\gamma = 0.1$, while long dashes mark Holt values for $\alpha = 0.1$ and $\gamma = 0.9$.

Gardner&McKenzie suggested to replace the forecast from Holt's method by a discounted forecast

$$\hat{x}_N(h) = L_N + \left(\sum_{j=1}^{h} \phi^j \right) T_N.$$

That forecast would be optimal for an ARIMA$(1,1,2)$ model of the form

$$\Delta X_t = \phi \Delta X_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}.$$

This model describes a variable that becomes stationary after first-order differencing, which may be more plausible for most economic series. This may motivate its use in economic forecasting. However, it is less often found in commercial computer software.

The true strength of the Holt-Winters procedure is that it allows for two variants of seasonal cycles. This is important for the prediction of quarterly and monthly economic series, which usually have strong seasonal variation. In the multiplicative version, level L_t, trend T_t, and seasonal S_t obey the recursions

$$\begin{aligned} L_t &= \alpha \frac{x_t}{S_{t-s}} + (1-\alpha)\left(L_{t-1} + T_{t-1}\right), \\ T_t &= \beta \left(L_t - L_{t-1}\right) + (1-\beta) T_{t-1}, \\ S_t &= \gamma \frac{x_t}{L_t} + (1-\gamma) S_{t-s}, \end{aligned}$$

where s is the period of the seasonal cycle, i.e. $s = 4$ for quarterly and $s = 12$ for monthly observations. The algorithm needs three parameters α, β, γ. Forecasts are generated according to

$$\hat{x}_N(k) = \left(L_N + T_N k\right) S_{N+k-s},$$

where S_{N+k-s} is replaced by the last available corresponding seasonal if $k > s$.

In the additive version, the recursions change to

$$\begin{aligned} L_t &= \alpha \left(x_t - S_{t-s}\right) + (1-\alpha)\left(L_{t-1} + T_{t-1}\right), \\ T_t &= \beta \left(L_t - L_{t-1}\right) + (1-\beta) T_{t-1}, \\ S_t &= \gamma \left(x_t - L_t\right) + (1-\gamma) S_{t-s}. \end{aligned}$$

Example. Figure 4 demonstrates the application of Holt-Winters filtering to the quarterly Austrian gross domestic product (GDP). In this clearly seasonal and trending series, both the multiplicative and the additive versions yield similar performance. None of them can, of course, predict the global recession of 2008.

The time-series program ITSM includes an additive version of the seasonal Holt-Winters algorithm with automatically determined smoothing parameters. For the Austrian GDP data, the algorithm determines $\alpha = 0.66$, $\beta = 0.05$, $\gamma = 1.00$. The value of γ implies that the algorithm does not find a separate seasonal component in the data and extrapolates any of the four quarters, simply based on the trend and noise components.

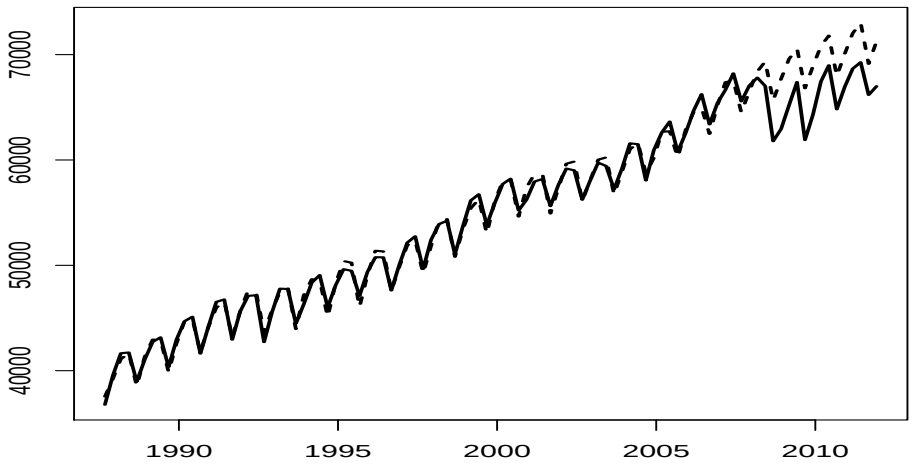

Figure 4: Austrian gross domestic product 1988–2012 and *Holt-Winters seasonal filtering*. Forecasts from 2008.

2.3 Brockwell & Davis' small-trends method

Among the many other *ad hoc* methods that have been suggested in the literature, the small-trends method by Brockwell&Davis deserves to be mentioned. It has been designed for variables with a 'not too large' trend component and a seasonal component.

B&D assume that the trend is not too strong, therefore they conclude that the average of observations over one year yields a crude indicator of the trend for that year. That 'trend', which is constant for any particular year, can be subtracted from the original observations. From the thus adjusted 'non-trending' series, averages for one particular period (month k or quarter k) are calculated from the whole sample. For example, a January component is calculated over all de-trended January observations. The resulting seasonal cycle and trend are then extrapolated into the future.

While there are obvious drawbacks of this method—the seasonal cycle is time-constant, no attempt is made at modelling the possibly stationary remainder component nor the trend itself—the method has proved to be quite reliable for short time series. In their original contribution, B&D mention a data set on monthly observations of road accidents for a time range of a decade. Ten observations are not enough to permit sensible time-series

modelling.

Example. An application of the small-trends method to the Austrian GDP series is shown in Figure 5. At the end of the sample, 16 observations were forecast from an extrapolation of the seasonal and trend components. As expected, the downturn of the global recession cannot be predicted perfectly.

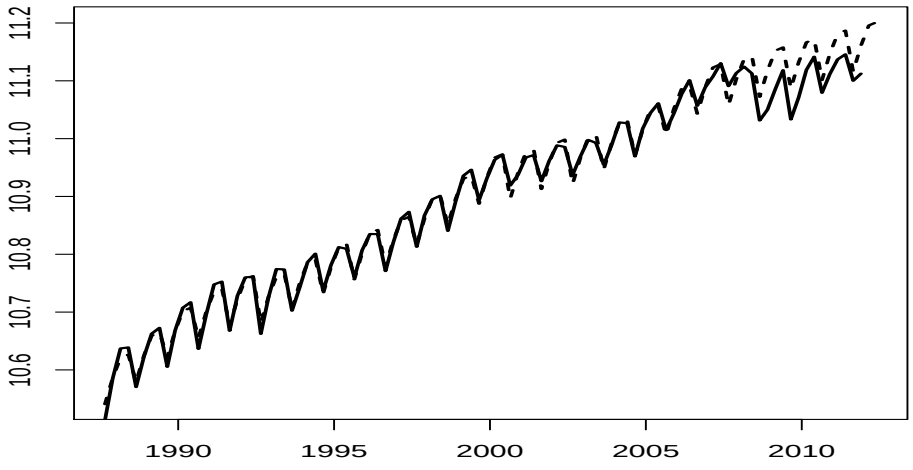

Figure 5: Austrian gross domestic product (solid curve), smoothed series, and four years of predictions from 2009 based on the small-trends method of Brockwell&Davis.

3 Univariate time-series models

Forecasts based on time-series models require some tentative specification of a statistical model that is conceivable as a data-generating process. At least for forecasting, it is not required that one believes that the used time-series model actually *did generate* the observations. Note that, particularly in the tradition of Box&Jenkins, models are used when they are *good* and discarded when they are *bad*. The quotation from G.E.P. Box

> *All models are wrong but some are useful.*

has become famous. It is obvious from the examples in the Box&Jenkins book that some of the suggested time-series models are unlikely to have generated the data. For example, Box&Jenkins use a second-order integrated ARIMA model for an airline passenger series. It is hard to imagine the implied quadratic time trend to persist in the future development of airline traffic, even in the most optimistic outlook for an airline carrier. Box&Jenkins stick to a philosophy of fitting models to observed data, as guided by some visual inspection of time plots, correlograms, and other similar means, with the purpose of short-term out-of-sample prediction. It is useful to review the basic time-series models that are available as candidates for modelling in the following subsections.

Prediction based on most of these models proceeds in an almost identical fashion. Most models have the general form

$$X_t = g\left(X_{t-1}, X_{t-2}, \ldots; \theta\right) + \varepsilon_t,$$

where $g\left(.\right)$ is a (non-linear or linear) function that depends on an unknown parameter θ and ε_t is unpredictable from the past of X_t. The most rigorous assumption would be that ε_t is *i.i.d.*, i.e. independent and identically distributed. Particularly in forecasting, the more liberal assumption of a *martingale difference sequence* (MDS) generally suffices for models of the above type. A process (ε_t) is said to follow a MDS if and only if

$$\mathrm{E}(\varepsilon_t | \mathcal{I}_{t-1}) = 0,$$

where $\mathcal{I}_{t-1}$ denotes some information set containing the process past. Additionally, it may be convenient to assume existence of some moments, for example a variance should at least exist. In any case, the MDS assumption is stronger than the backdrop assumption of linear time-series analysis, the *white noise*, defined by ε_t being uncorrelated over time. The white-noise assumption does not preclude predictability using nonlinear models, which would invalidate some of the arguments in forecasting.

After specifying a time-series model (class), a *model selection* stage follows in order to determine discrete specification parameters, such as lag orders. The selected model is *estimated* on the basis of a maximum-likelihood procedure or some convenient approximation thereof. The estimated parameters $\hat{\theta}$ are inserted in place of the true ones θ and the approximate conditional expectation

$$g\left(X_{t-1}, X_{t-2}, \ldots; \hat{\theta}\right) \approx g\left(X_{t-1}, X_{t-2}, \ldots; \theta\right) = \mathrm{E}\left(X_t | X_{t-1}, X_{t-2}, \ldots\right)$$

serves to determine $\hat{X}_{t-1}(1)$. The same mechanism can be extended to obtain multi-step predictions. For example, $g\left(\hat{X}_{t-1}(1), X_{t-1}, \ldots; \hat{\theta}\right)$ yields a two-step forecast.

Only in those cases where the design of the generating law is more complex, such as

$$X_t = g\left(X_{t-1}, X_{t-2}, \ldots; \varepsilon_t; \theta\right)$$

with genuinely non-linear dependence on the noise term, conditional expectation differs from $g\left(X_{t-1}, X_{t-2}, \ldots; 0; \theta\right)$ and it may pay to use stochastic prediction. For stochastic prediction, a statistical distribution is assumed for ε_t, and a large quantity of replications of the right-hand side $g\left(X_{t-1}, X_{t-2}, \ldots; \varepsilon_t; \hat{\theta}\right)$ are generated. The average of the replications then approximates the conditional expectation. As an alternative to generating ε_t from a parametric class of distributions, one may also draw ε_t replications from estimation residuals ('bootstrapping').

In nonlinear models of the latter type, ε_t should be independent over time and have some time-constant moments. The more liberal MDS assumption is less natural for these general models.

3.1 Linear models with rational lag functions

Most applications of time-series modelling use linear models. There are three basic types of linear models: autoregressive (AR), moving-average (MA), and ARMA models. One view on these models is that they provide good first-order approximations to the dynamics of the data-generating process. Another view is that they only intend to capture the first two moments of that DGP, i.e. means and covariances (including variance and correlations). Because all features are described exhaustively by the first two moments in a world of Gaussian distributions, the linear models are perfect for data that are nearly Gaussian or for samples that are too small that there would be evidence to the contrary. To the forecaster, these simple linear models are attractive as long as non-Gaussian or non-linear features are not strong enough

that they can be successfully exploited by different models. Even for many data sets that are known to contain non-linear dynamics, the non-linearity is either not time-constant enough or not significant enough to permit improved prediction. This justifies the popularity of these simple linear models, as they were suggested by BOX&JENKINS.

3.1.1 Autoregressive (AR) models

AR models are the most popular time-series models, as they can be fully estimated and tested within the framework of least-squares regression. A series X_t is said to follow an AR model if

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

This is the AR(p) model or the autoregressive model of order p. The error ε_t is usually specified as white noise, i.e. as uncorrelated over time with a constant variance and mean zero. Sometimes, time independence is also required. Time-series statisticians often prefer a notation such as Z_t for the white-noise error instead of ε_t. It is convenient to write the model in *lag operators,* defined as $BX_t = X_{t-1}$, as

$$\begin{aligned} X_t &= \phi_1 B X_t + \phi_2 B^2 X_t + \ldots + \phi_p B^p X_t + \varepsilon_t, \\ (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) X_t &= \varepsilon_t, \\ \phi(B) X_t &= \varepsilon_t. \end{aligned}$$

Some authors use L ($l$ag) instead of B ($b$ackshift), without any change in the meaning. The advantage of $\phi(B)$ lies in the isomorphism of lag polynomials and complex real-coefficients polynomials $\phi(z)$, with regard to many properties. For example, stability of the autoregressive model can be checked easily by calculating the roots (zeros) of $\phi(z)$. If all zeros of $\phi(z)$ are larger than one in absolute value, there is a stationary process X_t, which satisfies the autoregressive equation and can be represented as

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

The coefficients ψ_j converge to zero, such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Note that $\psi_0 = 1$. If some roots are less than one, there may be a stationary 'solution', but it is *anticipative*, i.e. X_t depends on future ε_t. Such a solution is not interesting for practical purposes. If some roots are exactly one in their modulus, no stationary solution exists. A typical case is the *random walk*

$$\begin{aligned} X_t &= X_{t-1} + \varepsilon_t, \\ (1 - B) X_t &= \varepsilon_t. \end{aligned}$$

There is no stationary process that satisfies this equation.

The coefficient series ψ_j is a complicated function of the p coefficients ϕ_j. For the simple AR(1) model, it is easy to determine ψ_j, as

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

is immediately transformed into

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

The AR(1) model admits a stationary solution—is 'stable'—whenever $|\phi| < 1$. For higher-order models, stability conditions on coefficients become increasingly complex. There is no way around an evaluation of the polynomial roots.

Stationary autoregressive processes have an autocorrelation function (ACF) $\rho_j = \text{corr}\,(X_t, X_{t-j})$ that converges to zero at a geometric rate as $j \to \infty$. For the AR(1) process, $\rho_j = \phi^j$ is very simple. Higher-order AR processes have cyclical fluctuations and other features in their ACF, before they finally peter out for large j. One hopes that the sample ACF (the *correlogram*) reflects this geometric 'decay', although it is of course subject to small-sample aberrations.

Another characteristic feature of AR(p) models is that the partial autocorrelation function defined as

$$PACF\,(j) = \text{corr}\,(X_t, X_{t-j} | X_{t-1}, \ldots, X_{t-j+1})$$

becomes exactly zero for values larger than p. Time-series analysts say that the PACF 'cuts off' or 'breaks off' at p. Again, the sample PACF (or *partial correlogram*) may reflect this feature by becoming 'insignificant' after p.

Prediction on the basis of an AR(p) model is easy, as one simply replaces the true coefficients by in-sample estimates $\hat{\phi}_j$ and thus obtains one-step forecasts immediately by

$$\hat{X}_{t-1}\,(1) = \hat{\phi}_1 X_{t-1} + \hat{\phi}_2 X_{t-2} + \ldots + \hat{\phi}_p X_{t-p}.$$

Similarly, the next forecast at two steps is obtained by replacing the unknown observation X_t by its prediction $\hat{X}_{t-1}\,(1)$

$$\hat{X}_{t-1}\,(2) = \hat{\phi}_1 \hat{X}_{t-1}\,(1) + \hat{\phi}_2 X_{t-1} + \ldots + \hat{\phi}_p X_{t-p+1}.$$

This algorithm can be continued to longer horizons. In that case, predictions will converge eventually to the long-run mean. In these simple specifications,

the long-run mean must be zero. However, AR models are often specified with a constant term

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

in which case the mean will be

$$\mathrm{E}X_t = \frac{\mu}{1 - \phi_1 - \ldots - \phi_p} = \frac{\mu}{\phi(1)}.$$

Note that the case $\phi(1) \leq 0$ has been excluded.

3.1.2 Moving-average (MA) models

A series X_t is said to follow a *moving-average* process of order q or MA(q) process if

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q},$$

where ε_t is again white noise. MA(q) models immediately define stationary processes, every MA process of finite order is stationary. In order to preserve a unique representation, usually the requirement is imposed that all zeros of $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ are equal or larger than one in modulus. In that case, the MA model corresponds to the Wold representation of the stationary process. Wold's Theorem tells that every stationary process can be decomposed into a deterministic part and a purely stochastic part, with the stochastic part being an infinite-order MA process. The errors (or *innovations*) of the Wold representation are defined as *prediction errors*, if X_t is predicted linearly from its past. This theorem motivates the general usage of MA models in linear time series.

If all zeros of $\theta(z)$ are larger than one in modulus, the MA process has an autoregressive representation of generally infinite order $\sum_{j=0}^{\infty} \psi_j X_{t-j} = \varepsilon_t$ with $\sum |\psi_j| < \infty$. This excludes stationary MA processes with unit roots in $\theta(z)$. Thus, MA models are, as it were, more general than AR models. MA processes with an infinite-order autoregressive representation are said to be *invertible*.

A characteristic feature of MA processes is that their ACF $\rho(j)$ becomes zero after $j = q$. In fact, it is fairly easy to calculate all values of $\rho(j)$ from given MA coefficients θ_j. The property of the ACF should be reflected in the correlogram, which should 'cut off' after q. By contrast, the partial ACF converges to zero geometrically.

Forecasting from an MA model requires estimating the coefficients θ_j from the sample, after identification of the lag order q. While Box&Jenkins suggested using visual inspection of correlograms for identifying q, most researchers nowadays use information criteria for this purpose. Estimates $\hat{\theta}_j$ are

obtained from maximizing the likelihood or from some approximating computer algorithm. Additionally, estimates of the errors ε_t must be obtained, again using some computer algorithm. If $\hat{\theta}_j$ and $\hat{\varepsilon}_s$, $s < t$ are available, a one-step forecast is calculated as

$$\hat{X}_{t-1}(1) = \hat{\theta}_1\hat{\varepsilon}_{t-1} + \hat{\theta}_2\hat{\varepsilon}_{t-2} + \ldots + \hat{\theta}_q\hat{\varepsilon}_{t-q}.$$

More generally, j–step forecasts obey

$$\hat{X}_t(j) = \sum_{k=j}^{q} \hat{\theta}_k\hat{\varepsilon}_{t+j-k}$$

for $j \leq q$, while forecasts will become trivially zero for $j > q$ steps.

3.1.3 Autoregressive moving-average (ARMA) models

A series X_t is said to follow an *autoregressive moving-average* process of order (p, q) or ARMA(p, q) process if

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

with ε_t white noise. The ARMA model is stable if all zeros of $\phi(z)$ are larger than one. The representation is unique if all zeros of $\theta(z)$ are larger or equal to one in modulus *and* if $\phi(z)$ and $\theta(z)$ do not have common zeros. The stable ARMA model always has an infinite-order MA representation. If all zeros of $\theta(z)$ are larger than one, it also has an infinite-order AR representation.

In practice, ARMA models often permit to represent an observed time series with a lesser number of parameters than AR or MA models. A representation with the minimum number of free parameters is often called *parsimonious*. Particularly for forecasting, parsimonious models are attractive, as the sampling variation in parameter estimates may adversely affect prediction. In small samples, under-specified ARMA models—i.e., with some parameters set to zero, while they are indeed different from zero—often give better predictions than correctly specified ones.

For ARMA processes, both ACF and partial ACF approach zero at a geometric rate. It is difficult to determine lag orders p and q for ARMA models from visual inspection of correlograms and partial correlograms. While some authors did suggest extensions of the correlogram (*extended ACF, extended sample ACF*), most researchers determine lag order by comparing a set of tentatively estimated models via information criteria.

After estimating coefficients $\hat{\theta}_j$ and $\hat{\phi}_j$ by some approximation to maximum likelihood and calculating approximate errors $\hat{\varepsilon}_s$, $s < t$, one-step forecasts are defined by

$$\hat{X}_{t-1}(1) = \hat{\phi}_1 X_{t-1} + \ldots + \hat{\phi}_p X_{t-p} + \hat{\theta}_1 \hat{\varepsilon}_{t-1} + \hat{\theta}_2 \hat{\varepsilon}_{t-2} + \ldots + \hat{\theta}_q \hat{\varepsilon}_{t-q}.$$

Two-step forecasts are obtained from

$$\hat{X}_{t-1}(2) = \hat{\phi}_1 \hat{X}_{t-1}(1) + \hat{\phi}_2 X_{t-1} + \ldots + \hat{\phi}_p X_{t-p+1} + \hat{\theta}_2 \hat{\varepsilon}_{t-1} + \ldots + \hat{\theta}_q \hat{\varepsilon}_{t-q+1}.$$

3.2 Integrated models

In many data sets, stationarity appears to be violated, either because of growth trends, regular cyclical fluctuations, volatility changes, or level shifts. In all these cases, expectation and/or variance appear to be changing through time. However, signs of non-stationarity do not necessarily justify discarding the linear models. In order to 'beat' ARMA prediction, a model class has to be found that can do better. In the absence of such a model class, it is advisable to work with the available ARMA structures. For the very special cases of trending behavior and of seasonal cycles, integrated models are attractive. The idea of an integrated process is that the observed variable becomes stationary ARMA after some preliminary transformations, such as first or seasonal differences.

In detail, a variable is said to be *first-order integrated* or **I(1)**, if it is not stationary while its first difference $\Delta X_t = X_t - X_{t-1}$ is stationary and invertible (i.e. excluding the case of unit roots in the MA part) ARMA. It is said to be *second-order integrated* or **I(2)**, if its second-order differences $\Delta^2 X_t = \Delta(X_t - X_{t-1}) = X_t - 2X_{t-1} + X_{t-2}$ are stationary and invertible ARMA. When the differences are ARMA(p, q), Box&Jenkins were using the notation ARIMA$(p, 1, q)$. Again, the 'I' in ARIMA stands for 'integrated'.

For the decision on whether to take first differences of the original variable, Box&Jenkins suggested visual analysis. The correlogram of a first-order integrated variable decays very slowly, while the correlogram of its first difference replicates the familiar ARMA patterns. The correlogram of second differences of an I(1) variable or of first differences of a stationary ARMA (or **I(0)**) variable is dominated by some strong negative correlations and tends to be more complex again. Box&Jenkins called this the case of 'over-differencing'. Following the statistical test procedures developed by Dickey&Fuller around 1980, most researchers opine that visual analysis may not be reliable enough. We note, however, that Box&Jenkins did not target statistical decisions and true models. For forecasting, differencing may imply an increase in precision, even when the original variable is

I(0) and *not* I(1). Motivated by robustness toward structural breaks and similar features, the benefits of 'over-differencing' were analyzed in detail by Clements&Hendry (1999).

The test by Dickey&Fuller requires running the regression

$$\Delta X_t = \mu + \tau t + \phi X_{t-1} + \gamma_1 \Delta X_{t-1} + \ldots + \gamma_p \Delta X_{t-p} + \varepsilon_t.$$

Significance points for the t–statistic of the coefficient ϕ were tabulated by Dickey&Fuller. The null hypothesis is I(1) or 'taking first differences'. If the t–statistic is less than the tabulated points, one may continue with the original data. The number of included lags p is usually selected automatically according to information criteria. For data without a clearly recognizable trend, the term τt is omitted. That version of the test requires slightly different significance points. In order to test for second-order differencing, the test procedure can be repeated for ΔX instead of X, with $\Delta^2 X$ replacing ΔX. This second-order test usually is conducted on the version without the term τt.

Economic variables are often sampled at a quarterly frequency. Such variables tend to show seasonal cycles that change over time yet have some persistent regularity. For example, construction investment may show a trough in the winter quarter, while consumption may peak in the quarter before Christmas. Then, the mean is not time-constant and the variables are not stationary. A traditional tactic is to remove seasonality by *seasonal adjustment* and to treat the adjusted series as an integrated or stationary process. Seasonal adjustment is performed by sophisticated filters that, unfortunately, tend to remove some dynamic information and to imply a deterioration in predictive performance.

Box&Jenkins suggested the alternative strategy of *seasonal differences*. Later authors called a variable *seasonally integrated* if its seasonal differences $\Delta_4 X_t = X_t - X_{t-4}$ are stationary and invertible ARMA. While Box&Jenkins based the decision on whether to apply seasonal differencing on a visual inspection of correlograms, Hylleberg *et al.* developed a statistical test, whose construction principle is similar to the Dickey-Fuller test. Note that seasonal differencing removes seasonal cycles *and* the trend, such that further differencing of $\Delta_4 X_t$ is usually not required on purely statistical grounds. By contrast, Box&Jenkins liberally conducted further differencing, which again is possibly justified if one aims at good prediction models rather than at true models.

In detail, Box&Jenkins suggested SARIMA models (seasonal ARIMA) with the notation $(P, D, Q) \times (p, d, q)$, which is to mean that D times the filter Δ_4 and d times the filter Δ is applied on the original series, while P

and Q stand for seasonal ARMA operators of the form $\Theta(B) = 1 + \Theta_1 B^4 + \Theta_2 B^8 + \ldots + \Theta_P \Theta^{4P}$ etc. Instances are rare, where the SARIMA approach appears to be appropriate. No example is known where $D > 1$ could be applied sensibly.

Additionally to ARMA modelling of seasonal differences, SARIMA models, and seasonal adjustment, a monograph by Franses offers another alternative. Franses considers *periodic models* that essentially have coefficients that change over the annual cycle, i.e. each quarter has different coefficients. Clearly, a drawback is these models require estimating four times as many parameters, which implies that reliable modelling needs longer samples.

At this point, we can proof the ARIMA(0,1,1) equivalence of SES. If the generating process is

$$\Delta X_t = \varepsilon_t + (\alpha - 1)\,\varepsilon_{t-1},$$

conditional expectation yields the forecast

$$\hat{X}_t(1) = X_t + (\alpha - 1)\varepsilon_t,$$

assuming we know the error ε_t at time point t. In theory, this ε_t can be retrieved as the difference $X_t - \hat{X}_{t-1}(1)$, and substitution immediately yields the recurrence equation for SES. In practice, α will not be known, and forecasts will differ from true conditional expectations.□

3.3 Fractional models

Fractional or *long-memory* models are linear models outside of the ARMA framework. While it is easy to define Δ^0 as the identity operator (i.e., nothing changes) and $\Delta^1 = \Delta$, it is less straightforward to define Δ^d for $0 < d < 1$. This can be done by expanding $(1 - z)^d$ as a power series

$$(1 - z)^d = 1 + \delta_1 z + \delta_2 z^2 + \ldots \tag{1}$$

It turns out that δ_j are given by binomial expressions. If the power series converges, one may define

$$\Delta^d X_t = X_t + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \ldots \tag{2}$$

Time-series models may assume that Δ^d is the required operator that yields a white noise or a stationary ARMA variable. Then, X_t can be called I(d) in analogy to the cases $d = 0, 1, 2, \ldots$ It can be shown that I(d) variables with $d < 0.5$ are stationary, while $d \geq 0.5$ defines a non-stationary variable. These fractional models describe processes whose autocorrelation function decays at a slower than geometric pace. They are sometimes used for variables with

some indication of 'persistent' behavior in conjunction with some signs of mean reversion, such as interest rates or inflation.

A drawback of these models is that they need very long time series samples for a reliable identification of d and of the ARMA coefficients of $\Delta^d X$. Therefore, they pose a particular challenge to the forecaster.

3.4 Non-linear time-series models

General non-linear time-series models have become popular through the book by TONG. A more recent and very accessible introduction was provided by FRANSES&VANDIJK. FAN & YAO and TERÄSVIRTA *et al.* are representative for the current state of the art in this field. Because non-linear structures require large data sets for their correct identification and for potential improvements in forecasting accuracy, the literature focuses on applications in finance and in other sciences, where large data sets are available.

ARCH (*autoregressive conditional heteroskedasticity*) models are a special case. Introduced by ENGLE in 1982 for monthly inflation data, their main field of applications quickly moved to financial series, for which various modifications and extensions of the basic model were developed. Unlike other non-linear time-series models, ARCH models have been rarely applied outside economics. Many researchers in finance apparently prefer *stochastic volatility* (SV) models to ARCH, which however are difficult to estimate.

3.4.1 Threshold autoregressions

The general non-linear autoregressive model

$$X_t = f\left(X_{t-1}, X_{t-2}, \ldots, X_{t-p}\right) + \varepsilon_t$$

for some non-linear function $f\left(.\right)$ would require non-parametric identification of its dynamic characteristics. Often, the identified function will not conform to stability conditions, such that simulated versions of the model tend to explode. A well-behaved and simple specification is the piecewise linear threshold model

$$X_t = \phi_{(j)} X_{t-1} + \varepsilon_t, \quad r_{j-1} < X_{t-1} < r_j, \quad j = 1, \ldots, k,$$

where $r_k = \infty$ and $r_0 = -\infty$. The literature has called this model the SETAR (*self-exciting threshold autoregressive*) model and has reserved the simpler name TAR for cases where $\phi_{(j)}$ may be determined from exogenous forces rather than X_{t-1}. The model can be—and has been indeed—generalized into many directions, such as longer lags, moving-average terms etc. SETAR

can be shown to be stable for strict white noise ε_t and for $\left|\phi_{(j)}\right| < 1$ for all j. It is interesting that this is only a sufficient condition. SETAR models can be stable for 'explosive' coefficients $\left|\phi_{(j)}\right| > 1$ in the central regions, i.e. $j \neq 1, k$. Economics applications include interest rates and exchange rates, which are series where thresholds can be given an interpretation.

Note that, while for given r_j all coefficients can be simply estimated by least squares, identification of k and of the r_k thresholds is quite difficult. If some 'regimes' are poorly populated in the sample, neither their exact boundaries nor their coefficients can be estimated reliably. Therefore, instances of successful forecasting applications are rare.

3.4.2 Smooth transition

The rough and maybe implausible change in behavior at the threshold points in the SETAR model has motivated the interest in smoothing the edges. Franses&VanDijk give a simple example of a *smooth-transition autoregressive* (STAR) model by

$$X_t = \phi_{(1)} X_{t-1} \left(1 - G\left(X_{t-1}; \gamma, c\right)\right) + \phi_{(2)} X_{t-1} G\left(X_{t-1}; \gamma, c\right) + \varepsilon_t,$$

where $G(.)$ is a transition function whose shape is determined by the parameters γ ('smoothness') and c ('center'). The left limit of the transition function should be 0 (for $X_{t-1} \to -\infty$) and the right limit should be 1 (for $X_{t-1} \to \infty$). Curiously, these transition functions—a typical choice is a logistic function—resemble the squashing functions of neural networks. The correct specification and identification of the functions is difficult. Like all time-series models, applications usually include deterministic intercepts that may also change across regimes.

3.4.3 Random coefficients

Instinctively, many empirical researchers think that the model

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

would be more general than a standard AR(1) and may respond to the needs of an ever-changing environment. Unfortunately, this model is void without specifying the evolution of the coefficient series ϕ_t. Stable autoregressions ($\phi_t = \psi \phi_{t-1} + \eta_t$) and white noise have been suggested in the literature. Curiously, some variants have been shown to come close to ARCH models. If $\mathrm{E}\phi_t = 1$, the model is sometimes said to contain a 'stochastic unit root'.

3.4.4 ARCH models

Many financial time series, such as returns on common stocks, are nearly unpredictable. Neither ARMA models nor non-linear time-series models allow reliable and possibly profitable predictions. However, these series digress in two aspects from usual white noise, as it would be generated from a Gaussian random number generator. Firstly, the unconditional distribution is severely leptokurtic, with more unusually large and small realizations than would be implied from the Gaussian law. Secondly, calm and volatile episodes are observed, such that at least the variance appears to be predictable. The ARCH model by Engle appears to match both features. A variable X is said to follow an ARCH process if it obeys two equations, which are called the 'mean equation' and the variance or volatility equation. For example, consider the simple ARCH(1) model

$$\begin{aligned} X_t &= \mu + \varepsilon_t, \\ \mathrm{E}\left(\varepsilon_t^2 | \varepsilon_{t-1}\right) &= h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2. \end{aligned}$$

The model is stable (asymptotically stationary) for $\alpha_0 > 0$ and $\alpha_1 \in (0,1)$. Later, it was found that it is actually also stable for some $\alpha_1 \geq 1$, depending on distributional assumptions. In these cases, the variable X_t has infinite variance but it is strictly stationary. The model yields leptokurtic X_t and it also implies sizeable 'volatility clustering', meaning that h_t is serially (positively) correlated. In stationary ARCH processes with finite variance, the mean and variance of X_t are time-constant. It is only the conditional variance h_t that is time-changing. If the mean equation is replaced by a stable ARMA $\Phi(B) X_t = \mu + \Theta(B) \varepsilon_t$, the generated variable X_t *is* ARMA. In this respect, the model is still 'linear'. Some authors even considered specifying a non-linear mean equation.

Today, the most popular variant of the ARCH model is the GARCH(1,1) model

$$\begin{aligned} X_t &= \mu + \varepsilon_t, \\ \mathrm{E}\left(\varepsilon_t^2 | \varepsilon_{t-1}\right) &= h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta h_{t-1}. \end{aligned}$$

The model may be extended by including a non-trivial first equation, for example an ARMA model. In the GARCH model, the dependence of h_t and h_{t-1} is modelled directly. Stationarity and finite variance are implied by $\alpha_0 > 0$, $\alpha_1 + \beta \in (0,1)$, $\alpha_1 > 0$, $\beta \geq 0$. Again, larger coefficients may retain stationarity, though not finite variances.

Judging by statistical significance of coefficients, the GARCH model was applied successfully to all kinds of financial variables, such as stock returns,

exchange rates, or interest rates, when these are observed at a moderately high frequency, such as weekly or daily. Prediction of volatility based on the models has been less successful. While most forecast evaluation criteria are tuned to mean prediction, it is doubtful whether prediction of X_t^2 is a good measure for the predictive accuracy of volatility forecasts. TAYLOR suggested various alternatives for this purpose, which have however not been taken up much in the literature. CLEMENTS (2005) provides a more recent review of this problem.

Many econometric computer programs now contain GARCH estimation routines. A slightly inefficient iterative routine for the ARCH specification was suggested by ENGLE in his original contribution. He suggests estimating both the mean equation and the variance equation by weighted least squares, replacing the $\hat{\varepsilon}_t^2$ for the h_t and determining the weights from the most recent iteration. Therefore, given a conveniently long sample, ARCH models can even be estimated by ordinary least-squares routines.

3.5 Neural networks

Based on the reports of some surprising successes of forecasts, neural network modelling has become popular in the 1990s. To econometricians, a stumbling block to the application of neural networks (NN) is the quaint language that was adopted by NN adherents. The book by CHATFIELD (section 3.4.4) allows an interesting insight into NN modelling and re-positions it among the other non-linear procedures.

The main difference to non-linear time-series modelling is the focus on unobserved variables, which serve as black-box connections between inputs and outputs. Similar unobserved 'layers' are used in the unobserved-components (UC) models suggested by HARVEY who calls them 'structural'. Note that unobserved variables also appear in ARCH models (h_t). Input variables, hidden variables, and output variables are connected by linear and S–shaped 'squashing' functions, the latter ones resembling those in the STAR models. These squashing functions motivate the word 'neural', as nerve cells or neurons react to a stimulus in a similar S–shaped way. Model selection and parameter estimation can be performed on the first portion of the data, which is then called the *'training set'*. NN adherents call the model selection stage the choice of *'architecture'*. Alternatively, the model selection and fitting may be done by optimizing the forecasting performance on a second part of the sample, the *'testing set'*. Implied parameter estimates are usually neither tabulated nor published. In this regard, NN modelling resembles Bayesian time-series applications such as BVAR (*Bayesian vector autoregression*, due to DOAN&LITTERMAN&SIMS). The suggestion to revise the architecture

and parameter estimates, as new data come in, is not specific to NN. Therefore, NN models 'learn' as much as other time-series models.

Apparently, the reported benefits of NN modelling with respect to prediction are mainly due to their usage of non-linear squashing functions. Most NN applications in economics use just one layer of hidden 'nodes'. CHATFIELD gives an example for the implied output reaction function of a single-layer NN, which here is simplified slightly to

$$\hat{x}_t = \phi_0 \left(w_{c0} + \sum_{h=1}^{H} w_{h0} \phi_h \left(w_{ch} + \sum_{j=1}^{h} w_{jh} x_{t-j} \right) \right),$$

with $\phi_j, j = 0, \ldots, h$ denoting functions and w_{jk} denoting coefficient parameters. Autoregressive linear combinations of lagged x with different maximum order cause a reaction by the unobserved hidden H 'neurons' that is described by the functions ϕ_h. A linear combination of the H neurons then causes another, potentially non-linear, reaction ϕ_0 in the systematic part of the x variable at time t. The 'stochastic' difference of $\hat{x}_t$ and x_t is not modelled explicitly.

3.6 State-space modelling

Like ARMA modelling and neural networks, state-space modelling is a different approach to time-series analysis rather than a class of different models. Often, state-space modelling leads to time-series models that are equivalent to ARMA models. A basic idea is that the observed series are generated as a 'contaminated version' of an output from an unobserved black-box system with its own intrinsic dynamic behavior. In other words, dynamic modelling is shifted from the observed variable to one or more unobserved variables.

In CHATFIELD's notation, an *observation equation*

$$X_t = h_t' \theta_t + n_t$$

explains the observed X as a linear combination (the vector h_t') of unobserved *state variables* θ_t plus a 'noise' error n_t. The *transition equation*

$$\theta_t = G_t \theta_{t-1} + w_t$$

extrapolates the unobserved system variables from their past. G_t is a square matrix that conforms to the dimension of the unobserved state, i.e. the number of state variables. Another stochastic error w_t is permitted. In this most general form, the system is clearly not identifiable. Under certain restrictions, estimates of h, θ_t, G etc. are obtained from an algorithm called

the *Kalman filter*. All ARMA models can be re-written in this form, with constant G and h.

The most popular specification for a state-space model in economics is the 'structural' unobserved-components model by HARVEY. Its simplest variant is

$$\begin{aligned} X_t &= \mu_t + n_t, \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + w_{1,t}, \\ \beta_t &= \beta_{t-1} + w_{2,t}. \end{aligned}$$

Note that the state is two-dimensional and has the variables μ and β. While β_t is a random walk, μ_t is an I(2) process. Therefore, X_t is implicitly also assumed as an I(2) process. Because we know that very few economic variables are likely to be I(2), the model is difficult to justify within the framework of linear modelling. The argument of HARVEY and his numerous followers is that the variances and relative variances of the errors n and w_1, w_2 make the model quite flexible in the sense that the variance of w_2 simply approaches zero if X_t is really an I(1) variable. Indeed, note that for $w_{2t} = 0$, X_t reduces to a random walk plus noise. On the other hand, it is not to be expected that the choice of three variance parameters allows an equally powerful modelling tool as $p + q$ ARMA parameters would. The reported success of HARVEY's models relies mainly on the fact that many economic series approximately correspond to the given dynamic structure. Whenever the dynamics digress from the basic pattern, the models can have a very poor performance.

It is shown easily that the above model corresponds to a specific ARIMA structure with $d = 2$, and that the ARMA coefficients are simple functions of the relative error variances. From these properties, it is quite clear that the model is not a generalization of I(1) models that allow the average growth rate to be time-changing, as some authors apparently believe. An advantage of the 'structural' models is that they directly yield estimates of 'local' trends μ_t and of 'local' growth rates β_t. The local trends provide smoothed versions of the original series that are sometimes interpreted as having economic contents. For example, a smoothed output series may be interpreted as potential output.

There is specialized computer software for 'structural' modelling, which also allows for some additional complexities, such as seasonal cycles in the state variables or even business cycles. It is also possible to allow for some autocorrelation in the noise terms and to estimate their intrinsic dynamics by non-parametric methods. That approach highlights the conceptual difference between ARMA and UC modelling. While ARMA modelling targets a description of the dynamic behavior of the observed variables, UC modelling

focuses on a decomposition of the observed series into a systematic and a noise part. The implied models as well as the underlying dynamic behavior of the observed variables can be the same in both approaches.

4 Multivariate forecasting methods

Usually, multivariate forecasting methods rely on *models* in the statistical sense of the word, though there have been some attempts at generalizing extrapolation methods to the multivariate case. This does not necessarily imply that these methods rely on *models* in the economic sense of the word. Rather, one may classify multivariate methods with regard to whether they are *atheoretical*, such as time-series models, or *structural* and theory-based. Words can be confusing, however, as the so-called 'structural time-series methods' suggested by HARVEY are actually atheoretical and not theory-based. Truly structural forecasting models with an economic background will be left for the next unit.

4.1 Is multivariate better than univariate?

Multivariate methods are very important in economics and much less so in other applications of forecasting. In standard textbooks on time-series analysis, multivariate extensions are given a marginal position only. Empirical examples outside economics are rare. Exceptions are data sets with a predator-prey background, such as the notorious data on the population of the Canadian lynx and the snowshoe hare. By contrast, the multivariate view is central in economics, where single variables are traditionally viewed in the context of relationships to other variables. Contrary to other disciplines, economists may even reject the idea of univariate time-series modelling on grounds of the theoretical interdependence, which appears to be an exaggerated position.

In forecasting, and even in economics, multivariate models are not necessarily better than univariate ones. While multivariate models are convenient in modelling interesting interdependencies and achieve a better (not worse) fit within a given sample, it is often found that univariate methods outperform multivariate methods *out of sample*. Among others, one may name as possible reasons:

1. Multivariate models have more parameters than univariate ones. Every additional parameter is an *unknown* quantity and has to be estimated. This estimation brings in an additional source of error due to sampling variation.

2. The number of potential candidates for multivariate models exceeds its univariate counterpart. Model selection is therefore more complex and lengthier and more susceptible to errors, which then affect prediction.

3. It is difficult to generalize nonlinear procedures to the multivariate case. Generally, multivariate models must have a simpler structure than univariate ones, to overcome the additional complexity that is imposed by being multivariate. For example, while a researcher may use a nonlinear model for univariate data, she may refrain from using the multivariate counterpart or such a generalization may not have been developed. Then, multivariate models will miss the nonlinearities that are handled properly by the univariate models.

4. Outliers can have a more serious effect on multivariate than on univariate forecasts. Moreover, it is easier to spot and control outliers in the univariate context.

An additional complication is *conditional forecasting*, which means that a variable Y is predicted, while values for a different variable X are assumed over the prediction interval. If X is a policy variable, it may make sense to regard it as fixed and not to forecast it. Even then, true-life future X may react to future values of Y, while that reaction is ignored in the conditional forecast. The arguments of 'feedback' by CHATFIELD and the so-called Lucas critique are closely related. Important contributions on the feedback problem, with some relation to forecasting, are due to GRANGER (Granger causality) and to ENGLE&HENDRY&RICHARD (weak and strong exogeneity). If X is merely another variable that is also predicted over the prediction interval, though maybe using a simple univariate extrapolation method, these forecasts may be wrong and there will be another source of error.

If the forecasting model is designed such that forecasts of one or more variables of type X are not generated, while one or more variables of type Y are modelled as being dependent on X, the system is called *open-loop*. If all variables are modelled as dependent on each other and on lags, the system is called *closed-loop*. Even closed-loop systems may allow for deterministic terms, such as constants, trend, or other variables that can be assumed as known without error at any future time point.

4.2 Static and dynamic forecasts

A clear distinction of *in-sample model fitting* and *out-of-sample forecasting* is even more important than for univariate models. If $1, \ldots, T$ serves as the time range for a 'training set', a model is selected on the basis of these observations, the model parameters are estimated from the very same set, and predictions are then generated for $2, \ldots, T$ according to the generation

mechanism of the model, the 'prediction errors' will be mere *residuals*. If variables are predicted over the time range $T+1, \ldots, T+h$ without using any further information on the 'test set', the differences between observations (if available) and predicted values will be true out-of-sample prediction errors. The exercise yields a sequence of a one-step, a two-step, ..., and an h–step prediction, which are sometimes called a *dynamic forecast*. There are many intermediate cases.

In *static forecasts*, true observations are used for all lagged values of variables. This situation does *not* correspond to a forecaster who only uses information on the training set. Prediction errors from static forecasts are difficult to classify. One would expect that they tend to increase from $T+1$ to $T+h$ if the forecasting model misses an important part of the DGP, which results in the impression of a time-changing DGP. Otherwise, the errors can be regarded as one-step out-of-sample prediction errors. If observations are used for explanatory current variables in regression-type prediction models, the resulting error comes closer to a residual. The expression *ex-post prediction* is used for both variants. Sometimes it is also used for mere in-sample fitting and calculating residuals.

Other intermediate cases are obtained if, for example, a model is selected on the basis of the total sample $1, \ldots, T+h$, while parameters are estimated from $1, \ldots, T$ only. The resulting forecast is not really out-of-sample, and its prediction error tends to under-estimate the true errors. Such exercises are often presented in research work and are sometimes labelled incorrectly as out-of-sample prediction.

4.3 Cross correlations

An important exploratory tool for modelling multivariate time series is the *cross correlation function* (CCF). The CCF generalizes the ACF to the multivariate case. Thus, its main purpose is to find linear dynamic relationships in time series data that have been generated from stationary processes.

In analogy to the univariate case, a multivariate process X_t is called (covariance) stationary if

1. $\mathrm{E}X_t = \mu$, i.e. the mean is a time-constant n–vector;
2. $\mathrm{E}\left(X_t - \mu\right)\left(X_t - \mu\right)' = \Sigma$, i.e. the variance is a time-constant positive definite $n \times n$–matrix Σ;
3. $\mathrm{E}\left(X_t - \mu\right)\left(X_{t+k} - \mu\right)' = \Gamma\left(k\right)$, i.e. the covariances over time depend on the lag only, with non-symmetric $n \times n$–matrices $\Gamma\left(k\right)$.

The matrix-valued function of $k \in \mathbb{Z} : k \mapsto P(k)$, which assigns correlation matrices instead of covariance matrices, is then called the CCF. One may note that $p_{ij}(k) = p_{ji}(-k)$, such that also the CCF is informative for non-negative k only or, alternatively, for positive and negative k and $i \leq j$. For estimates of the CCF in computer programs, the latter version is usually preferred. Definitions vary across programs and authors, such that $P(k)$ may be equivalent to somebody else's $P(-k)$. This is important insofar as a predominance of, for example, negative values $p_{ij}(-k)$ over the positive values $p_{ij}(k)$ with regard to their size and significance is often taken as indicating a causal direction from the j–th variable to the i–th variable etc. This is a crude check, anyway, as Granger causality can be assessed reliably only from the CCF of pre-filtered values or from multivariate autoregressive or, preferably, moving-average representations.

Identification of vector autoregressive (VAR) or other multivariate models from the empirical CCF is difficult. Also the issue of significance bounds for the CCF values is tricky. Automatically generated bounds from computer packages only serve as rough guidelines. The main basic result is Bartlett's formula, which is given here for the case of a bivariate process ($n = 2$):

$$\begin{aligned}\lim_{T\to\infty} T\mathrm{cov}\left(\hat{p}_{12}(h), \hat{p}_{12}(k)\right) &= \sum_{j=-\infty}^{\infty} [p_{11}(j)\, p_{22}(j+k-h) + p_{12}(j+k)\, p_{21}(j-h) \\ &\quad -p_{12}(h)\left\{p_{11}(j)\, p_{21}(j+k) + p_{22}(j)\, p_{21}(j-k)\right\} \\ &\quad -p_{12}(k)\left\{p_{11}(j)\, p_{21}(j+h) + p_{22}(j)\, p_{21}(j-k)\right\} \\ &\quad +p_{12}(h)\, p_{12}(k)\left\{\frac{1}{2}p_{11}^2(j) + p_{12}^2(j) + \frac{1}{2}p_{22}^2(j)\right\}]\end{aligned}$$

This extremely complicated formula simplifies for low-order vector MA processes and, of course, for white noise. For example, if one of the two processes is white noise and the true $p_{12}(h) = 0$ for all but finitely many h, then the asymptotic variance of $\hat{p}_{12}(h)$ will be 1 for these h. As for the ACF and correlograms, a very slow decay of the CCF indicates non-stationarity of the integrated or unit-root type. Even if integrated variables are indicated, taking first differences is not necessarily recommended. The problem of whether to apply differences in a multivariate setting has led to the extended literature on *co-integration*.

Extensions of the usual cross-correlogram analysis are the cross-correlogram analysis of pre-whitened variables and phase plots. If variables are individually pre-whitened, the diagonal elements of the CCF matrix will be trivial, while the off-diagonal elements may be more representative of the true dynamic interaction among variables. For discrete-time variables, a *phase plot*

refers to a scatter plot of a variable X_t and a lag, such as X_{t-1} or X_{t-j}. If X_t has been generated by a first-order autoregression $X_t = \phi X_{t-1} + \varepsilon_t$, points should show a straight line with the slope corresponding to ϕ. Such phase plots are particularly valuable if a nonlinear time-series relationship is suspected. With multiple time series, also phase plots of X_{it} versus $X_{j,t-k}$ can be considered, although the number of possible diagrams becomes large and the variety of diagrams can become confusing.

4.4 Single-equation models

Two types of so-called single-equation models can be considered for multivariate forecasting: regression models and transfer-function models. Both types are 'open-loop' models and model a dynamic relationship of an 'endogenous' variable that depends on one or several 'explanatory' variables. The methods are helpful in forecasting only if future values of the explanatory variables are known, if forecasting of explanatory variables is particularly easy, or at least if there is no dynamic feedback from the endogenous to the explanatory variables. In the latter case, the condition is similar to the requirement that there is *no Granger causality* running from the endogenous to the explanatory variables and that there is *Granger causality* from the explanatory to the endogenous variables. The concepts may not be entirely equivalent, as Granger causality often only refers to linear relationships and additional conditions may be imposed by the parameterization of interest. These issues have been investigated in the literature on *exogeneity* (Engle&Hendry&Richard).

The linear *regression model* can be assumed as known. In a time-series setting, errors are often correlated over time, such that ordinary least squares (OLS) may not be appropriate and some kind of generalized least squares (GLS) procedure should be considered. These GLS models assume two equations, with a time-series model for the errors from the 'means' equation, for example an ARMA model:

$$\begin{aligned} y_t &= \beta' x_t + u_t \\ \phi(B)\, u_t &= \theta(B)\, \varepsilon_t. \end{aligned}$$

The specification $\theta(B) = 1$ and $\phi(B) = 1 - \phi B$ yields the classical GLS model that is used in the Cochrane-Orcutt or Prais-Winsten estimation procedures. When these models are applied for forecasting, one does not only need predicted values for x but also for u. Because u_t is unobserved, it is usually replaced by in-sample residuals from the means equation and by out-of-sample predictions based on these residuals. The difference between true errors u and estimated residuals $\hat{u}$ brings in another source of uncertainty,

adding to the sampling variation in the coefficient estimates $\hat{\beta}$, $\hat{\phi}$, and $\hat{\theta}$. In many cases, it is more advisable to replace the means equation by a dynamic version that contains lagged values of y_t and of some x_t as further explanatory variables, and whose errors are approximately white noise (*dynamic regression*). An efficient search for the optimal dynamic specification of such equations can be lengthy and may resemble the techniques used in transfer-function modelling. Thus, the two methods are not too different in spirit. It can be shown that static GLS models are a very restrictive version of dynamic regression models. Some authors recommend removing trends and seasonals from all variables or possible differencing before regression modelling, which however may be problematic and may entail a deterioration of predictive accuracy.

Transfer-function models are based on dynamic relationships among mean- and possibly trend-corrected variables y_t (output) and x_t (input) of the form

$$\delta(B)\, y_t = \omega(B)\, x_t + \theta(B)\, \varepsilon_t. \tag{3}$$

Three different lag polynomials must be determined, which can be difficult. A common suggestion is to start from an ARMA or ARIMA model for the input variable x_t and to apply the identified 'filter' to both input and output. It is then much easier to determine the remaining polynomial. Of special interest is the case that $\omega_j = 0$ for $j < d$. This indicates that x affects y with a *delay* and that, for any h–step forecast for y_t with $h \leq d$, forecasts for the input x_t are not required. In this case, x is also called a 'leading indicator' for y. In analogy to the GLS regression model, efficient forecasting needs predictions for the unobserved error series, which are constructed by extrapolating in-sample residuals.

4.5 Vector autoregressions and VARMA models

These models usually treat all n variables in a vector variable X as 'endogenous'. For stationary multivariate processes, a convenient parameterization is the vector ARMA (VARMA) model

$$\Phi(B)\, X_t = \Theta(B)\, \varepsilon_t \tag{4}$$

with the multivariate white noise series ε_t, defined by $\mathrm{E}\left(\varepsilon_t \varepsilon'_{t-k}\right) = 0$ if $k \neq 0$. The model is stable if all roots of $\det(\Phi(z)) = 0$ are larger than one in absolute value. The model is invertible, i.e. X can be expressed as a convergent infinite-order vector autoregression, if all roots of $\det(\Theta(z)) = 0$ are also larger than one. Uniqueness requires the roots of $\det(\Theta(z)) = 0$ to be larger or equal to one. Unfortunately, this condition is necessary but not

sufficient for a unique representation. This is one of the reasons why vector ARMA models are rarely used in practice.

For $\Theta(z) \equiv 1$, one has the vector AR or VAR model

$$\Phi(B) X_t = \varepsilon_t,$$

which is quite popular in economics. The simple VAR(1) model for the bivariate case $n = 2$ can be written as

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

with the 2×2–matrix

$$\Phi = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}.$$

Of special interest are cases such as $\phi_{12} = 0$. Then, the first variable X_1 does not depend on past values of the second variable, although it may be correlated with it and it may depend on its own past via ϕ_{11}. There is no Granger causality from X_2 to X_1. The model reduces to a transfer-function model. X_1 can be forecasted via the univariate model $\hat{X}_{1,t-1}(1) = \hat{\phi}_{11} X_{1,t-1}$ and generally as

$$\hat{X}_{1,t-1}(h) = \hat{\phi}_{11}^h X_{1,t-1}.$$

Building on this forecast, a forecast for X_2 can be constructed from the second equation. Similar remarks apply to the reverse case $\phi_{21} = 0$.

For the general VAR(1) case, h–step forecasts are obtained from

$$\hat{X}_{t-1}(h) = \hat{\Phi}^h X_{t-1},$$

where $\Phi^2 = \Phi\Phi$, $\Phi^3 = \Phi^2\Phi$ etc. Similarly, forecasts from VAR(p) models can be constructed in strict analogy to the univariate case. The forecasts require coefficient estimates and a lag order p, which are determined from the 'training' sample ending in $t-1$. While cross correlograms may be helpful in finding a good lag order p, many researchers choose it by minimizing multivariate information criteria. Estimation can then be conducted by OLS, which can be shown to be efficient in the absence of further coefficient restrictions.

VARX and VARMAX models are extensions of the VAR and VARMA framework, which allow for exogenous ('X') variables whose dynamics are not specified or whose dynamics at least does not depend on the modelled 'endogenous' variables. For forecasting, the X variables require an extrapolation technique or assumptions on their future behavior.

4.6 Cointegration

Although VAR modelling traditionally assumes stationarity of all series, it is not generally recommended to difference non-stationary components individually, as such a step may destroy important dynamic information. The critical issue is *cointegration*. In short, when there is cointegration, cointegrated models should be used for forecasting. When there is no cointegration, series with an integrated appearance should be differenced.

Cointegration is said to exist in a vector time series X_t when some of the components are individually first-order integrated (I(1)) while the remaining components may be stationary, and there is a linear combination of all components expressed by a vector β such that $\beta' X$ is stationary (I(0)). Cointegration is non-trivial when β has non-zero entries at the integrated components. Apart from this *CI(1,1) cointegration*, which reduces integration order from one to zero, higher-order cointegration can be defined but its practical applicability is limited and examples for applications are rare. Note that there are slightly different definitions of CI(1,1) cointegration in the literature but that the above definition is the most convenient for multivariate VAR modelling.

The issues at stake can be best motivated by considering a VAR system that possibly contains some integrated and some stationary components

$$\Phi(B) X_t = \varepsilon_t.$$

It is easily shown that such systems can always be transformed into an *error-correction form*

$$\Delta X_t = \Phi X_{t-1} + \Gamma_1 \Delta X_{t-1} + \ldots + \Gamma_{p-1} \Delta X_{t-p+1} + \varepsilon_t, \tag{5}$$

where the matrices Φ and $\Gamma_1, \ldots, \Gamma_{p-1}$ are one-one functions of the autoregressive coefficient matrices $\Phi_1, \ldots, \Phi_p$. Note that the lag order of the polynomial $\Gamma(z) = I - \Gamma_1 z - \ldots - \Gamma_{p-1} z^{p-1}$ is $p-1$, while the lag order of $\Phi(z)$ is p. The crucial matrix is Φ. If the rank of Φ is zero, there is no cointegration and the system is best handled in differences. If the rank of Φ is n, the system is stable and should be handled in the original form without differencing. If the rank is between 1 and $p-1$, there is cointegration. In this case, the rank of Φ yields the number of linearly independent cointegrating vectors in the system. Empirically, the rank of Φ is determined by a sequence of tests with the null hypothesis 'rank $\Phi \leq k$' and the alternative 'rank $\Phi > k$'. All details are found in JOHANSEN's monograph. This rank determination procedure has been integrated into many econometric software packages, such as RATS and EViews.

Once the rank has been determined, estimation of (5) is conducted by *reduced-rank regression*, a technique that was developed by T.W. Anderson. Usually, this estimation procedure has been implemented in econometric software together with the rank-determination sequence. Then, estimates $\hat{\Phi}$ and $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_{p-1}$ can be substituted for the true matrices and h–step forecasts are retrieved easily.

Ignoring cointegration issues and naively differencing non-stationary components implicitly assumes the restriction $\Phi = 0$. If the true $\Phi \neq 0$, this does not only imply *mis-specification* due to an incorrect restriction but this mis-specification has particular consequences. The forecasts will not reflect the property that $\beta' X$ is stationary, in contrast with the sample. In economics, the usual interpretation is that 'equilibrium conditions are ignored'. For example, consumption or investment may amount to a relatively constant share of national income in the sample, while that share may rise or fall in the forecast interval. Forecasts may violate theory-based plausibility conditions. Using simulation and algebraic techniques, the consequences of cointegration for forecasting were analyzed by Clements&Hendry and in two influential studies by Engle&Yoo and by Christoffersen&Diebold. While Engle&Yoo demonstrated that cointegration may only be important for prediction at longer horizons, such as $h > 10$, Christoffersen&Diebold showed that such results depend on which variable is targeted in prediction. Accurate prediction of the original components may be a different task from predicting stationary variables such as ΔX and $\beta' X$. In some specifications, cointegration may even be important for one-step prediction, while, in other specifications, even long-horizon predictions are best conducted on the basis of the original 'level' VAR $\Phi(B) X_t = \varepsilon_t$, while the differences VAR $\Gamma(B) \Delta X_t = \varepsilon_t$ will typically suffer from serious mis-specification. As a bottom line, it is recommended to use cointegration techniques and the implied prediction models. When no cointegration is found, one should use VARs in differences.

4.7 Multivariate state-space models

State-space models, such as Harvey's 'structural' time-series models, can be generalized to the multivariate case. Chatfield draws particular attention to the SUTSE (*seemingly unrelated time-series equations*) model with common trends

$$\begin{aligned} X_t &= \Theta \mu_t + \mu^* + n_t, \\ \mu_t &= \mu_{t-1} + w_t. \end{aligned}$$

In this model, the matrix Θ is assumed as rectangular $n \times r$ with $r < n$, such that the second equation updates an r–dimensional integrated process. Some restrictions on μ^* and Θ are needed to identify the model. It can be shown that $n-r$ cointegrating relationships are implied by this model. The random-walk process μ_t is viewed as the 'common trend' of the system, which yields one out of several similar though conflicting definitions of 'common trends'. In strict analogy to the remarks on univariate state-space models, also these models are not really generalizations of VAR or vector ARIMA models and they do not really allow for time-varying structure. They are equivalent to certain vector ARIMA models and merely use a different parameterization.

5 Forecasts using econometric models

Even today, the basic workhorse tool for forecasting in economics is the large structural econometric model. These models are developed in specialized institutions, government agencies, and banks. They often consist of hundreds of equations. It is interesting that econometric theory has not been focusing on these models for almost thirty years. Following the general failure of these macro-econometric models to cope with the economies of the 1970s following the OPEC crisis, theory has discarded them (see Fair, for a noteworthy exception to the rule). This attitude was supported by several empirical studies that generally concluded that time-series models yield better forecasts than structural models. From the angle of economic as well as econometric theory, Sims recommended multivariate time-series modelling instead of structural models. He explicitly refers to the many 'incredible' exclusion restrictions of macro-econometric models, while he also concedes that 'large-scale models do perform useful forecasting ... functions'. Chatfield's assertion that 'econometric models are often of limited value to forecasters' sounds much harsher and may miss the point, as many large-scale models are built with the purpose of forecasting in mind. Recently, calibrated dynamic stochastic general-equilibrium (DSGE) models with a strong basis in economic theory but with sometimes only loose fit to data have been used widely in policy evaluation and long-range scenarios. The value of such models for short-run macroeconomic forecasting is still unsettled.

5.1 The classical macro-econometric model

The classical approach to macro-econometric modelling consists of several steps. Model building is indeed *guided by theory* but the model specification is not really *determined by theory*. I distinguish the following phases:

1. choosing the variables to be included in the model;
2. separating the variables into *endogenous* and *exogenous* variables;
3. sketching *a priori* causal relationships among the variables, in the style of a flow chart;
4. specification of estimable equations;
5. estimation;
6. forecasting.

Step 1 is guided by the intended purpose of the model. A model for the Austrian economy may include various Austrian economic variables and a lesser variety of foreign variables, together with possibly some non-economic variables such as population or weather conditions. An important restriction is the availability of data. For example, a quarterly model could include macroeconomic sub-aggregates of consumer demand or capital formation, while several government expenditure categories are available on an annual basis only. Mixing different periodicities is sometimes considered but many modelers find this option inconvenient.

Step 2 is crucial. Philosophers of science have named the classification of variables into endogenous and exogenous variables as a characteristic of econometric research (Cartwright). In many other sciences, an analogous classification rests on firm *a priori* beliefs, while such beliefs are rare in economics. No generally accepted theory can tell whether saving causes investment or investment causes saving. Similarly, there is no accepted causal ordering between the interest rate and real output. *Deterministic variables*, such as constants and time trends, are always *exogenous*. If the model is designed to represent the domestic sector of a small open economy, *foreign variables* are also usually regarded as *exogenous*. This may change if a world model is targeted. Typical foreign variables would be world demand and price indexes for the world market, while export prices and even import prices may respond to the domestic price structure. Many macroeconomic models also view *fiscal* variables as *exogenous*. This decision is often dictated by institutional reasons, such as the preference of a government institution who funds the forecasting project and does not want to be seen as a reaction node without a free will. One may, however, also choose that option to allow for conditional forecasting, i.e. scenarios that respond to alternative government policies. Some macroeconomic models choose an intermediate solution and keep aggregate tax revenues as endogenous, while they impose exogenous tax rates and long-range conditions such as a tendency of the budget to be balanced. *Non-economic variables* are usually modelled as exogenous.

Step 3 determines the internal structure of the model. Often, the large amount of considered variables is broken down into blocks of variables, such as real demand variables, price indexes, labor market variables etc. Influences across these blocks or sectors may run in both directions, though such relationships may be restricted to very specific effects. For example, the labor market may respond to the price sector via a Phillips-curve only. Causal orderings among blocks or variables are often not meant to specify dynamic causal orderings. Rather, they indicate which variables will appear on the right-hand side of regression equations for specific left-hand side variables. While the explanatory variables may be accepted as lags in further specifica-

tions, some model builders have a skeptical view on such time lags and may regard them as weak spots of their models. The economic science finds it difficult to specify dynamic interaction and has to justify it by arguments of habit persistence, adjustment costs, delivery lags etc. There is little information on the exact form of these features, hence coefficients that describe dynamic interaction are difficult to interpret. On the other hand, dynamic rather than static relationships in equations help in forecasting as well as in stabilizing the system. In the extreme case of an n–variable static model, a 'solution' for a given time point is calculated by solving a system of n often nonlinear equations. Forecasting will not be possible, unless assumptions are made on extrapolating $n-1$ of the variables at a different (future) time point. The other extreme is a VAR or VARX system, where all influences are dynamic. A forecast for future time points is calculated by simple insertion of the current observed values.

Step 4 consists in a cursory search for equation specifications. In most macro-econometric models, this search is conducted on a single-equation basis, in spite of persistent theoretical criticism. The dependent variable is specified, for example private consumption of durable goods, and various combinations of explanatory variables are tried as regressors. The specification search is guided by the set of possible influences that was specified in step 3. Functional forms, such as linear or double-log specifications, are selected according to statistical criteria or, less often, to theory arguments. Possible nonlinear forms in some equations are among the advantages of econometric modelling relative to linear VAR systems. Usually, the final specification is required to fulfil the following criteria:

1. regressors should be statistically significant;

2. coefficients should have an economically interpretable size and sign whenever such interpretation is possible;

3. influences that are deemed to be important for theoretical reasons should exist, which often implies 'keeping' insignificant regressors;

4. residuals should not show substantial autocorrelation;

5. if inserted into the whole model, the equation should not de-stabilize the model. A solution should be possible in a 'range' of values around the observed data and for a reasonable 'range' of possible residuals without obtaining inadmissible values, such as negative unemployment rates;

6. measures such as R^2 should be conveniently large, the required values ranging between 0.3 and 0.95;

7. more general statistical tests of mis-specification should not indicate severe violations of assumptions.

In many cases, these different targets are in conflict with each other. Modelers decide on their preferences among targets on an *ad-hoc* basis. It is no wonder that PINDYCK&RUBINFELD call model building a 'science and art'.

Step 5 demands for detailed comments. According to the classical model-building framework, as set out by the Cowles Commission, *linear* econometric models can be written in the form

$$Y\Gamma = XB + U.$$

Many econometric models are nonlinear but this form may even serve as a basis for estimating nonlinear structures. Y is a matrix of dimension $T \times G$ and contains the observations on the G 'endogenous' variables, while Γ is a $G \times G$–matrix of estimable coefficients. X is a matrix of dimension $T \times K$ and contains the 'exogenous' variables, *including lags of endogenous variables*. Due to their similar estimation properties, exogenous (including lagged ones) and lagged endogenous variables are summarized as *predetermined variables*. Note that the wording 'predetermined' does not refer to a forecasting application, as data on current exogenous variables are usually not available before the data on endogenous variables. B is a $K \times G$ coefficient matrix, while U contains regression errors.

Clearly, the matrix Γ cannot be estimated without further restrictions. It contains information on simultaneous relationships among the endogenous variables and represents the completely 'static' aspect of the model. Without restrictions, the system contains G different equations that look all alike. The extreme case $\Gamma = I$ is called the *reduced form* of the system. In a certain context, systems with $\Gamma = I$ are also called SUR models (*seemingly unrelated regressions*), as the only connection among equations is by common regressors and by possible correlation of errors, which two aspects may be ignored at first sight. Reduced-form systems can always be estimated by regression methods. They are convenient for forecasting but economists avoid them, as they do not reflect interdependencies that were developed in economic theory. Another argument against their usage is that their parameters are difficult to interpret. This argument again reflects the 'hybrid' aims of econometric modelling: even when the model is built exclusively for prediction, economists require its interpretability in economic terms.

The conditions on Γ that have to be imposed in order to make the model 'identified', i.e. to allow its sensible empirical estimation, are rather complex. Most applied econometric models fulfil these conditions anyway, due to their tendency to include 'many unbelievable restrictions', as SIMS has put it. In these systems with general Γ, regression estimation of single equations yields inconsistent parameter estimates. The literature recommends variants of instrumental variables estimation procedures, such as *two-stage least squares* (2SLS), in order to overcome this deficiency of least squares in 'simultaneous systems'. In empirical model building, this advice is not always followed. Many large-scale econometric models are estimated by least squares methods, while in other models 2SLS estimation is restricted to some sensitive sub-systems. An informal argument in favor of this practice is that the inconsistency of least squares plays a lesser role in larger models. This may not be true in general.

Finally, step 6 consists in substituting estimates for the unknown parameters and to predict the endogenous variables via

$$\hat{Y}_t(1) = X_{t+1}\hat{B}\hat{\Gamma}^{-1},$$

which poses difficulties if X_{t+1} is not available. Unless X contains only lagged endogenous variables—this is a special case of VAR modelling—X_{t+1} will not be available and has to be guessed or extrapolated. Other forecasters avoid this 'zero-residual' forecasting and prefer

$$\hat{Y}_t(1) = X_{t+1}\hat{B}\hat{\Gamma}^{-1} + \hat{U}_{t+1}\hat{\Gamma}^{-1},$$

where $\hat{U}_{t+1}$ are so-called 'add factors' that are guessed and inserted by the forecaster. This may make sense when the model has an annual frequency and the current year is to be forecasted. Although the endogenous variables are not yet available, some information has spread on the current economic situation, which may discourage the automatic usage of zero residuals. In practice, the calibration of add factors according to informal information is the principal and most time-consuming task of economic model forecasters. The above formulae are too simplistic, as they are valid for linear models only. With nonlinear models, simple multiplication by the inverted $\hat{\Gamma}$ is replaced by a numerical solution of a high-dimensional system of nonlinear equations. In former times, this step required a large amount of computer time, which was a binding constraint to forecasters. Nowadays, computers have become so powerful that even high-dimensional systems can be solved within a few seconds.

5.2 An example

A prototypical example for a macro-econometric model is the national income model by GRANGER:

$$\begin{aligned} C_t &= a_1 + b_1 Y_t + e_{Ct}, \\ I_t &= a_2 + c_2 P_{t-1} + e_{It}, \\ T_t &= d_3 GDP_t + e_{Tt}, \\ P_t &= a_4 + b_4 Y_t + f_4 I_{t-1} + e_{Pt}, \\ GDP_t &= C_t + I_t + G_t, \\ Y_t &= GDP_t - T_t. \end{aligned}$$

The *endogenous* variables are: consumption C, investment I, taxes T, profits P, gross domestic product GDP, disposable income Y. Apparently, the original GRANGER called Y the 'national disposable income' out of a misunderstanding, thus excluding the government sector from the national economy. Rather, Y appears to be a disposable income of the household and firm sectors. The *exogenous* variables are: government expenditure on goods and services G and the constant. There are four *structural* or behavioral equations with error terms and two identities. The economy is closed, which is clear from the GDP definition. There are two simultaneous feedback cycles: C depends on Y, Y depends on GDP, and GDP depends on C; T depends on GDP, GDP depends on C, C depends on Y, and Y depends on T. Therefore, least-squares estimation will yield inconsistent estimates of all parameters (coefficients).

By substitution, for example GDP can be expressed by the *predetermined* variables

$$GDP_t = A + \frac{G_t}{H} + c_2 \frac{P_{t-1}}{H} + e_{Gt},$$

where $A = (a_1 + a_2)/H$, $H = 1 - b_1 + b_1 d_3$, and $e_{Gt} = (e_{Ct} + e_{It} - b_1 e_{Tt})/H$. Analogous substitution yields the *reduced form* of the system, which expresses all six endogenous variables by predetermined variables and error terms. Note that the error structure of the reduced form necessarily has a singular covariance matrix. The equations of the reduced form can be estimated consistently by least squares. Econometric textbooks consider the option of retrieving estimates for the original 'structural' coefficients 'backward' from these reduced form estimates by algebraic operations and call it *indirect least squares*. This method is rarely used in practice.

In order to forecast GDP_{t+1}, one may use

$$\widehat{GDP}_t(1) = A + \frac{G_{t+1}}{H} + c_2 \frac{P_t}{H},$$

which assumes that approximations to the reduced-form coefficients A, H^{-1}, and c_2/H are available in any case, either from least squares on the reduced form or from two-stage least squares on the structural model and simple 'forward' algebraic operations. While P_t is available at t, G_{t+1} is not. Either one inserts a 'plausible value' for G_{t+1}, for example from the government's budget plan, or one uses an additional time-series or econometric equation for predicting G. The latter option 'essentially completes the system by treating G as if it were endogenous' (GRANGER), i.e. by endogenizing government policy. GRANGER suggests 'to use a model to forecast G and then to alter this forecast subjectively if relevant extra information is available to the forecaster. The success or lack thereof of such adjustments clearly depends on the quality of the information being utilized and the abilities of the forecaster who tries to use it'.

6 The LIMA forecasting model of the Institute for Advanced Studies Vienna

[Note: as a tool for professional short-term forecasting, the LIMA model is subject to a never-ceasing process of updates and minor modifications. Most details and parameter estimates in this section reflect the situation as of 2009.]

The LIMA model has grown out of the LINK project that attempts to join worldwide economic forecasting models into a common framework. Because many of the variables are only available at an annual frequency, the LIMA model also operates at this annual frequency. This can be troublesome for short-run prediction, as unofficial provisional data on main accounts aggregates come in on a quarterly basis. Therefore, LIMA is rarely run in its original form with zero residuals, and 'add factors' play a key role.

The LIMA model is a traditional macro-econometric prediction model with an emphasis on the economy's demand side. Thus, the model may be called a 'Keynesian' model. It has 99 equations, which implies 99 endogenous variables. As in most macro-econometric models, most equations are mere identities. Only 19 equations are 'behavioral' and contain estimated coefficients. With 99 endogenous variables and 19 structural equations, the LIMA model is a comparatively small macro-econometric model.

LIMA's model structure is updated frequently. Some equations may be replaced by better ones, while others are being eliminated in a search to simplify the overall structure and again others are being added in order to satisfy the needs for a more refined modelling of a certain sector of the economy. In the recent decade, the need for disaggregated modelling in many sectors has decreased on average. Accordingly, the 'true dimension' of the model has shrunk from a maximum of more than 30 to 19. The 'full dimension' was even larger, due to a very elaborate bookkeeping model for the Austrian public sector. Notwithstanding all modifications, the basic LIMA model structure still resembles its predecessors from the late 1970's.

Parameter estimates are updated once a year, when the official provisional data for the previous year become available, which is usually in September. For example, in September 2007 all equations were re-estimated using data from 1976 to 2006. For many equations, estimation intervals ended in 2005 if provisional data were earmarked as likely to be revised further. 1976 is the earliest year, for which national accounts data are available that correspond to the ESA standard. With some temporary exceptions, all equations are estimated by OLS. Indications of mis-specification due to autocorrelation are adjusted by dynamic modelling rather than by GLS–type corrections.

Thus, most behavioral equations are dynamic.

The model's center piece is the *domestic demand sector*. Demand aggregates are modelled in real terms, i.e. at constant prices, and sum up to real gross domestic product (GDP). Additional equations are used to determine *prices and deflators*. By multiplying those deflators into the real aggregates, nominal variables and eventually nominal gross domestic product (GDP$) are calculated. Following the transition from Paasche indexes to chained indexes in national accounts, adding-up identities of real variables do not hold exactly any more. This problem is resolved using artificial residual entities that are small on average and will be ignored in further discussion.

This adding-up to obtain GDP requires export and import variables. The treatment of exports and imports is asymmetric. *Imports* are fully endogenous and respond to demand categories, such as consumer durables and equipment investment. By contrast, exports are mainly exogenous. Older LIMA versions considered modelling exports as depending on world demand but, unfortunately, data on world demand become available with a considerable time lag only, which excludes its usage for the practical purpose of forecasting. For export and import prices, the approach is reversed. Import prices are exogenous, as it is assumed that Austrians have to accept the world market's price level, while export prices are endogenous.

Another component of GDP is public consumption. In the current version, public consumption is exogenous. In earlier versions, nominal public consumption was modelled as resulting from the sum of spending categories of general government. This practice was abandoned, as most government spending categories are exogenous and as the resulting price deflator of public consumption was often implausible or caused instabilities in model solution. In contrast to spending, several components of government revenues are modelled as endogenous variables, such as direct taxes or contributions to social security. From this *government sector*, balancing items such as the budget deficit can also be calculated.

The real and government sectors also interact with the *labor market sector*, which yields variables such as employment, the labor force, and wages. Other variables, such as the working-age population, are exogenous.

The LIMA model does not include a *financial sector*. Financial variables that are influential for the goods market, such as exchange rates and interest rates, are supplied by specialists on the financial sector who use separate models.

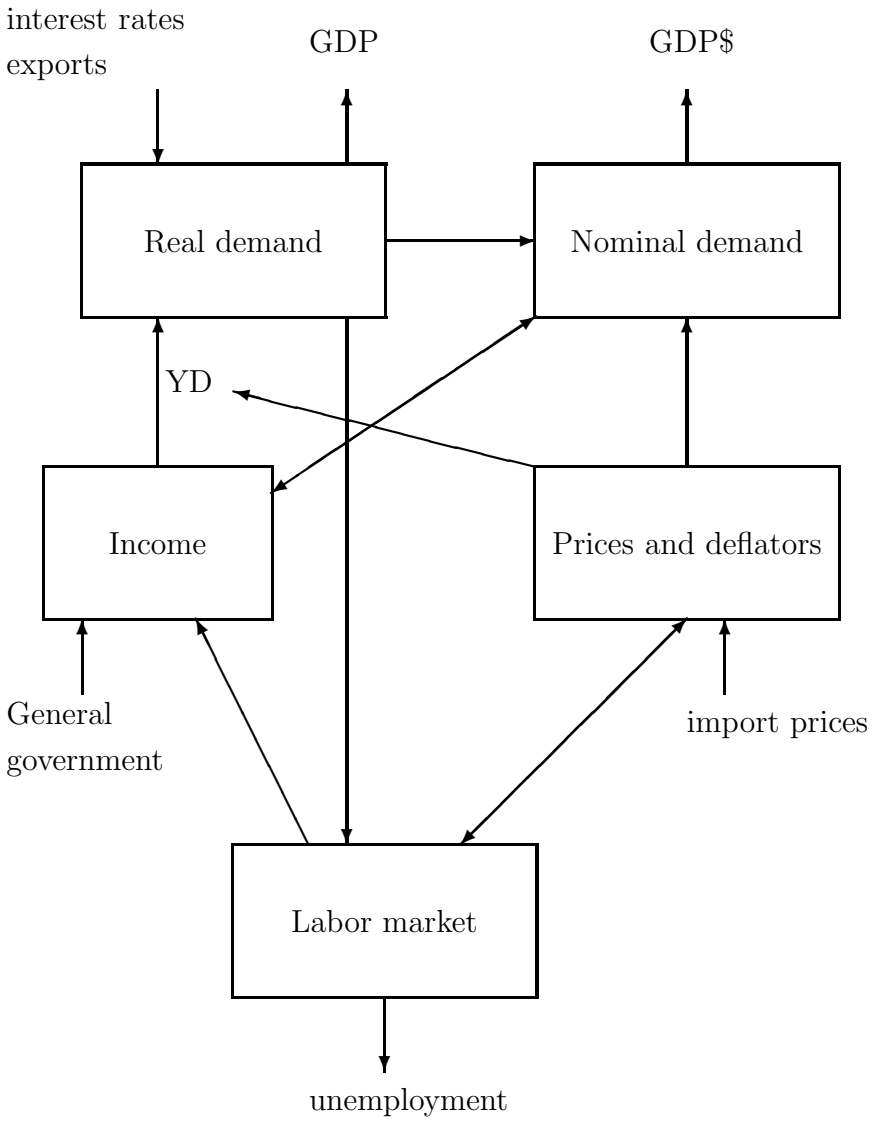

Figure 6: Structure of the forecasting model LIMA.

6.1 A typical demand equation: private consumption

Consumer demand consists of three categories: consumer durables, consumption of other goods, and consumer services. Almost 50% of household expenditures are spent on services. The share of services in household consumption appears to be increasing in the longer run. Before 1980, it used to be below 45%. Until a few years ago, consumer demand equations could be specified for the sub-aggregates, and total private consumption was obtained by adding the three forecasted categories. Unfortunately, updates of this decomposition are no more available with an acceptable time lag, and hence private consumer demand is now modelled on the basis of one behavioral equation only.

As a general rule, demand equations use logarithmic growth rates as dependent variables. Logarithmic growth rates are fairly constant in the longer run, hence they come closer to fulfilling the assumption of stationarity than, for example, first differences. On the other hand, percentage growth rates are far less convenient to handle from an econometric model builder's viewpoint.

In consumer demand, the principal explanatory variable is the growth rate of household disposable income yd. The real variable yd is obtained from deflating nominal household income by the consumption deflator. Thus, the same price index deflates income and the dependent variable.

Another potential source of explanation are error-correction relationships based on the 'great' consumption-output ratio. A cointegrating regression of log private consumption on log income

$$c_t = b_0 + b_1 yd_t + u_t$$

yields $\hat{b}_1 = 1.067$, slightly above one. One may argue, anyway, that unit elasticity for consumption should be imposed on the model, as otherwise longer-run forecasts may yield over-consumption and ever-decreasing saving rates. Therefore, the theory-based long-run restriction $b_1 = 1$ is used in further modelling.

The estimation results (see Table 1) are acceptable. The estimation time range has been shortened at its beginning, in order to exclude some outliers that were caused by changes in the VAT tax rate. Maybe excepting lagged income growth, all regressors are significant, and the (here, not very reliable) Durbin-Watson statistic does not indicate any serious specification error. The R^2 is lower than in comparable consumption equations for other data, which would be an incentive to search for further explanatory variables. Obvious suggestions, such as interest rates at any lags or lags of the dependent variable, failed to achieve significance. Several other variables were tried, for

example budget-consolidation indicators motivated by a Ricardian argument or unemployment indicators that may set saving incentives due to households' fears of getting unemployed. Again, none of these indicators achieved significance.

Apart from a local dummy for the exceptional year 1983, the equation shows a 'Keyenesian' short-run reaction with an elasticity of around 0.8 to real income, which is mitigated by a lesser contrary effect in the year afterwards. The error-correction term has the correct sign and has convincing significance. Austrian households tune their consumption patterns to longer-run saving plans and correct past under- or over-consumption.

Table 1: Behavioral equation for private consumption. Estimation time range is 1979–2006. Dependent variable is $\log(c_t/c_{t-1})$.

regressor	coefficient	t–value
constant	1.337	3.721
dummy83	0.042	4.924
$\log(c_{t-1}) - \log(yd_{t-1})$	-0.298	-3.712
$\Delta \log(yd_t)$	0.802	6.734
$\Delta \log(yd_{t-1})$	-0.205	-1.666

$R^2 = 0.809$, DW=1.710

6.2 Equipment investment

Besides consumption, investment or 'gross fixed capital formation' is another important component of aggregate demand. While the ESA system disaggregates investment into a larger number of subcomponents, LIMA only considers equipment investment, which includes machinery and transportation equipment, and construction investment, which includes business as well as residential construction. Equipment investment is the slightly smaller part but its equation is more important than the construction investment counterpart, as construction often relies much on public funding and policy.

Until fairly recently, modelling of investment demand relied on a theory-based factor-demand framework and on a CES or Cobb-Douglas production function. It was found, however, that dynamic error-correction modelling entailed better in-sample fit and also more stable out-of-sample predictive performance.

A cointegrating regression of logged equipment investment on gross domestic product (GDP) yields the form

$$\log(ife_t) = -4.09 + 1.323 \log(y_t) + ec_t,$$

where ife denotes equipment investment and y simply denotes GDP. The large value of the elasticity coefficient may sound a warning, as it appears to imply an ever-increasing share of equipment investment in output, while theory may tell that the investment-output ratio is approximately constant in the long run. Over the estimation sample, the relative increase of equipment investment was accommodated by a similar relative decrease of construction investment, such that the overall investment quota remained stable. For long-run projections, however, a different nonlinear framework may be more appropriate.

The residual from the cointegrating regression can be used as a regressor in an error-correction equation, with first differences of logarithmic investment as the dependent variable. It turns out that the influence of the error-correction term does not only have the correct sign, it is also clearly significant. Short-run influences are expressed via the regressor $\Delta \log(y_t)$ and a lag $\Delta \log(ife_{t-1})$, while further lags of these variables did not prove influential. The short-run effect of current real economic growth has a coefficient of 2.42, which confirms traditional wisdom that investment tends to over-react over the business cycle. The own lag is positive at 0.35 and reflects some persistence of investment decisions.

Table 2: Behavioral equation for equipment investment. Estimation time range is 1978–2006. Dependent variable is $\Delta \log(ife)$.

regressor	coefficient	t–value
dummies	~ -0.1	~ -2.8
ec_{t-1}	-0.651	-3.771
$\log(y_t/y_{t-1})$	2.417	4.716
$\Delta \log(ife_{t-1})$	0.347	2.422
real interest	-0.008	-2.258

$R^2 = 0.700$, DW=1.593

Of special interest is the effect of interest rates on investment. Economic theory suggests a negative influence from real interest rates on investment demand. Unfortunately, such an influence is often difficult to find in empirical

evidence. A lengthy search among various constructions for real interest rates resulted in a long-term ten-year rate that is deflated by investment prices. Its coefficient is significantly negative. However, it is to be noted that this influence is sensitive to the specification of the intercept. Ultimately, a version was preferred that excludes the insignificant intercept, excepting the constant from the cointegrating regression, and uses some time-local dummies for years with aberrant investment behavior. The usage of such dummies should be restricted to occasions where they are absolutely necessary, have convincing significance, and have an economic explanation, in this case usually due to changes in tax legislation.

6.3 An example for a deflator equation: investment prices

For each demand aggregate, two behavioral equations must be specified: an equation for real demand and an equation for the price deflator. In the case of equipment investment, the corresponding price deflator is named $pife$, for '$p$rices of $i$nvestment $f$ixed $e$quipment'. A large part of equipment investment demand is satisfied by imported goods, therefore the price deflator should be influenced directly by import prices. Another explanatory variable is ulc, unit labor costs, which stems from the labor market sector of the LIMA model. Substantial autocorrelation in the deflator also requires the insertion of a lag. Like many other model equations, the $pife$ equation is a dynamic regression equation. As a general rule, dynamic equations support the stability of the model, while static equations may result in unstable behavior. Therefore, in spite of some indication of remaining residual correlation, the $pife$ equation is satisfactory.

Table 3: Behavioral equation for the deflator of equipment investment. Estimation time range is 1978–2006. Dependent variable is $\Delta \log(pife_t)$.

regressor	coefficient	t–value
$\Delta \log(pife_{t-1})$	0.582	5.097
$\Delta \log(ulc_t)$	0.160	2.004
$\Delta \log(pmg_t)$	0.122	2.342
output gap	0.177	1.555

$R^2 = 0.738$, DW=2.480

The large value of the Durbin-Watson statistic should be seen against the backdrop of its tendency to be biased toward the ideal value of two in dynamic regression. Thus, the evidence on negative autocorrelation is stronger than would otherwise be indicated by a value of 2.48. Such negative residual correlation may point to an over-fit caused by too many regressors or it may be caused by the omission of the constant. It is to be noted that this equation, in line with most price equations, does not have a constant term. This implies that individual demand aggregates do not have an inflationary core of their own but that they just pick up price developments of their inputs.

6.4 Employment: no more Phillips curve

The employment equation ranges among the most frequently revised model equations. Earlier versions often included inflation among the regressors, while the present specification relies on error correction and on relative prices, i.e. the real wage ywr defined as nominal wage deflated by the GDP deflator. The main determinant of employment, however, is real output growth. The coefficient on real output growth shows the effects that are otherwise known as Okun's law.

Table 4: Behavioral equation for employment excluding self-employment. Estimation time range is 1981–2006. Dependent variable is $\Delta \log(le_t)$.

regressor	coefficient	t–value
constant	0.230	3.275
d83	-0.023	-4.003
$\Delta \log(y_t)$	0.450	5.091
$\log(le_{t-1}/y_{t-1})$	-0.174	3.401
$\log(ywr_{t-1})$	-0.221	-3.374

$R^2 = 0.704$, DW=2.070

All regressors are significant and have the expected signs. Unfortunately, the inclusion of a dummy variable was necessary. Fortunately, it is located in the earlier years and may have only small effects on forecasting.

The short-run Okun-type coefficient has the plausible value of 0.45. Note that it is not exactly the same as in Okun's law, due to some non-linear transformations and due to the omission of the labor-supply effects that are also captured in the original Okun coefficient. Error correction has a sizeable impact, which implies that the long-run unit elasticity shows its effects after few years already. In other words, a sudden recession has only small effects

on employment, while the full negative effects are felt if the recession does not end soon.

The negative effects of real wages, as the relative price of the production factor labor, are also quite strong. The variable *ywr* is the *per capita* gross wage divided by the GDP deflator. Technically, it counteracts the tendency of employment to grow in proportion to output, which would imply an absence of technological progress. However, the long-run growth of real wage puts a brake on unlimited employment expansion. Thus, the employment equation is a stabilizing component in the LIMA model, even though its structure will certainly have to be re-considered from time to time.

6.5 Theory and practice and the LIMA model

For many years, the LIMA model has proved a valuable backbone of the official IHS economic forecast. According to most comparative evaluations, the IHS economic forecast and that of its main competitor, the WIFO Institute for Economic Research, are of a comparable quality. Usually, forecasts are published for the current and for the following year only. Once a year, a medium-term projection is also presented to the public.

The following points can be identified where the LIMA forecast does not correspond to textbook forecasting:

1. There is no sharp boundary between a sampling interval and a prediction interval. Most macroeconomic variables—exceptions are exchange rates and unemployment data—spend years in an intermediate stage, where they are known with an increasing degree of precision.

2. Often, predicting the final data, i.e. those that mark the endpoint of all revisions by statistical agencies, is *not* targeted. For example, a forecast for 2000 may become uninteresting in 2003, even when it perfectly coincides with the final value.

3. The basis for prediction does not coincide with the estimation interval. For example, a forecast for 2007 is based on provisional data for all lagged variables from until 2006, while some model parameters may not have been updated from values beyond 2005.

4. Add factors play a key role. Incoming information is reflected in sizeable adjustment of residuals. Usually, zero-residuals forecasts are far off the mark.

5. Exogenous variables for the prediction interval are updated on an *ad hoc* basis. Some of these, however, correspond to information provided by other institutions—for example, government spending—or by researchers who use separate forecasting models.

6. All estimation is conducted by OLS, even when it is known that this procedure yields inconsistent estimates.

These points should not be interpreted as a critique of the current practice. Rather, one may assume that the discrepancies between the textbook prediction approach and current practice follow a longer-run experience of forecasters on how to do forecasting efficiently. It may be interesting to extend the textbook framework, taking the practitioners' approach into account. This direction of research is still in an early stage.

7 Evaluating predictive accuracy

The question "How good is a forecast?" comprises two separate aspects: firstly, measuring predictive accuracy *per se*; secondly, comparing various forecasting models. For example, if a variable is almost unpredictable, all forecasts are likely to be poor. Yet, a forecaster may still look for the best forecast among the poor ones.

The most commonly reported measures of predictive accuracy are

1. mean squared prediction errors or a variant of them;
2. mean absolute prediction errors;
3. percentage measures, such as the *mean absolute percentage error* (MAPE);
4. Theil coefficients;
5. significance measures, such as the DIEBOLD-MARIANO test statistic.

Additionally, some researchers, particularly in applied economics, use or suggest *qualitative* accuracy measures. CHATFIELD focuses on such a measure under the name of 'Percent Better'. We should scrutinize each of these suggestions in turn.

Note that the words *accuracy* and *precision* are not synonyms. Precision usually refers to the uncertainty of the forecast, for example measured by its variance, and does not imply accuracy.

7.1 Mean squared prediction errors

In the notation of CHATFIELD's textbook, the *prediction mean square error* PMSE is defined as

$$PMSE = m^{-1} \sum_{t=N-m+1}^{N} \left(x_t - \hat{x}_{t-1}(1)\right)^2 .$$

In this form, PMSE evaluates out-of-sample one-step errors. N observations are available, and the last m observations are used for evaluation. The forecast $\hat{x}_{t-1}(1)$ is meant to be calculated on the basis of the observations $x_1, \ldots, x_{t-1}$ only, including parameter estimates. The formula is easily modified for h–step errors with $h > 1$.

The PMSE formula does not specify m. Economic forecasters often tend to keep m/N small. For large m/N, the forecasts with small t are based on models estimated on short samples and thus be not representative. For small

m/N, however, the test sample may be too short to be a reliable indicator of absolute or relative accuracy. A potential modification would be to replace PMSE by a weighted MSE, with the weights increasing in t. This suggestion emphasizes the possibility that the true model may change over time. Then, the approximation by the prediction model toward the end of the sample is potentially more important for forecasts beyond N than the approximation in the earlier portion.

As $m \to \infty$, the PMSE should converge to a variance. Depending on the properties of the DGP, if such a one is assumed to exist, that variance should be close to—though slightly larger than—the variance of the theoretical prediction error $\mathrm{E}\left(x_t - \mathrm{E}\left(x_t|\mathcal{I}_{t-1}\right)\right)^2$, where $\mathcal{I}_{t-1}$ is an information set that contains the history of the series x. That theoretical variance serves as the benchmark for the construction of many statistical procedures, including least-squares estimation and AIC, which justifies the widespread usage of the PMSE.

The most common critiques of the PMSE are:

1. quadratic loss may not correspond to the forecaster's loss function;
2. the PMSE depends on scales;
3. the PMSE is vulnerable to outliers.

The idea of a forecaster's loss function is that forecast errors entail costs, in the sense that costs depend on

$$\mathrm{E}g\left(x_t - \hat{x}_{t-1}\left(1\right)\right)$$

for some function g. Some authors even consider generalizations of the expectation measure, as the costs may depend on time or on a nonlinear transformation of the vector of prediction errors, or on a more qualitative evaluation. The function should obey $g\left(0\right) = 0$ and $g\left(x\right) > g\left(y\right)$ for $x > y > 0$ and $x < y < 0$. Otherwise, g may be asymmetric and it may converge to a finite constant as its argument approaches infinity. The main problem with the loss-function approach is that the true loss or cost function is rarely known in practice. In some applications, even the existence of a cost function is uncertain, as the forecast may satisfy curiosity rather than serve as a basis for actual immediate decisions. Therefore, only simple *ad hoc* loss functions are considered usually, such as $g\left(x\right) = x^2$ and $g\left(x\right) = |x|$. The first choice yields the PMSE, the second one yields the mean absolute error, which is occasionally preferred due to its robustness toward outliers.

The scale dependence of PMSE is not a problem, as long as a specific variable x is in focus. If several variables are predicted using various procedures in each case, individual PMSEs should be weighted. A suggestion for weighting would be the sample variance of each series.

Often, instead of the PMSE, its square root is reported. While the PMSE corresponds to a measure of variance, the root MSE is a measure of standard deviation, which facilitates its interpretation. The transformation does not change the ranking of models and predictions according to their accuracy, and it also does not remove any of the inherent problems of the PMSE.

7.2 Mean absolute prediction errors

In the same notation as above, the *mean absolute error* is defined as

$$MAE = m^{-1} \sum_{t=N-m+1}^{N} |x_t - \hat{x}_{t-1}(1)| .$$

As $m \to \infty$, the MAE should converge to $\mathrm{E}\,|x_t - \mathrm{E}\,(x_t|\mathcal{I}_{t-1})|$ or a slightly larger value, assuming this value exists. For a Gaussian world, this moment is proportional to the standard deviation, with a fixed proportionality factor. Also for other distributions, the absolute moment will measure the dispersion of the forecast errors.

The MAE is based on the loss function $g(x) = |x|$, which is more sensitive to small deviations from 0 and much less sensitive to large deviations than the usual squared loss. Therefore, the MAE can be viewed as a 'robust' measure of predictive accuracy. The MAE tends to prefer forecasting procedures that produce occasional large forecast failures, while they are reasonably good on average. By contrast, the MSE tends to prefer forecasting procedures that avoid large forecast failures, even though they produce a less satisfactory fit otherwise.

Because the estimation procedures are usually based on least-squares criteria, an emphasis on the MAE may involve a slight logical inconsistency. The best class of models is then selected according to a criterion that is different from the one that selects among the different members of an individual model class.

7.3 Mean absolute percentage error

In the above notation, the *mean absolute percentage error* (MAPE) is defined by

$$MAPE = m^{-1} \sum_{t=N-m+1}^{N} \left| \frac{x_t - \hat{x}_{t-1}(1)}{x_t} \right|.$$

This definition answers complaints by some researchers that traditional criteria, such as PMSE and MAE, depend on the scaling of the variable x, which may be inconvenient if the criteria are used for comparing predictive accuracy across different variables or different time ranges. Unfortunately, the MAPE achieves scale independence by a simple division by x_t. This entails a serious drawback.

Many economic variables, such as stock returns and most growth rates, vary around zero. Whenever $x_t = 0$, the contribution at time point t and therefore the MAPE are undefined. Even if x_t is only approximately zero, the relative contribution of time point t will be enormous. Usually, there is no justification for preferring a high precision for small values of x_t.

It is obvious that the MAPE is tuned to variables that live in an area that is separated from zero by common sense. Economic examples would be the main aggregates of national accounts, such as fixed investment and private consumption.

Under the name of 'rmse percent error', Pindyck and Rubinfeld consider the direct squared-loss counterpart to the MAPE

$$m^{-1} \sum_{t=N-m+1}^{N} \frac{(x_t - \hat{x}_{t-1}(1))^2}{{x_t}^2}.$$

This measure has properties that are similar to those of the MAPE, with whom it shares most of its problems. While PMSE is more often reported than MAE, MAPE appears to be more popular than the above suggestion (see also Makridakis *et al.*).

In order to obtain a scale-free precision measure, it would be more appealing to consider measures such as

$$\frac{\sum_{t=N-m+1}^{N} (x_t - \hat{x}_{t-1}(1))^2}{\sum_{t=N-m+1}^{N} (x_t - \bar{x})^2}.$$

In this formula, the denominator measures the 'total' variation of x, while the numerator measures that part of the variation that has been accounted for by the prediction procedure. Thus, the measure is reminiscent of the regression R^2. Theil's accuracy measures follow a similar idea.

7.4 Theil coefficients

The idea of Theil's coefficients was to evaluate a forecast against the background of a simple or primitive forecast. If a forecasting procedure is to be taken seriously, it should at least 'beat' the simple benchmark. Unfortunately, it is not always clear which benchmark to use. THEIL used mainly random-walk or no-change forecasts, while other researchers use autoregressive prediction or exponential smoothers instead.

The version of Theil's coefficient that has been implemented into the EViews software is defined as

$$U = \frac{\sqrt{\sum_{t=N-m+1}^{N} \left(x_t - \hat{x}_{t-1}\left(1\right)\right)^2}}{\sqrt{\sum_{t=N-m+1}^{N} x_t^2} + \sqrt{\sum_{t=N-m+1}^{N} \hat{x}_{t-1}\left(1\right)^2}}.$$

For a 'good' predictor, the numerator will be small compared to the denominator. For a 'bad' predictor, both will be of similar magnitude. Theil's measures have often been criticized in the literature. They tend to yield implausible results in the sense that a predictor that optimizes them may have undesirable properties and *vice versa*, at least under lab conditions. It is not so certain whether this is also true in practical applications.

7.5 Decomposing the mean squared error

According to the EViews manual and to PINDYCK&RUBINFELD, the mean squared forecast error can be decomposed as

$$m^{-1} \sum \left(x_t - \hat{x}_{t-1}\left(1\right)\right)^2 = \left(\sum \hat{x}_{t-1}\left(1\right) - \bar{x}\right)^2 + \left(s_{\hat{x}} - s_x\right)^2 + 2\left(1 - r_{x\hat{x}}\right) s_{\hat{x}} s_x.$$

Here, $s_{\hat{x}}$ and s_x are sample standard deviations of $\hat{x}$ and x, respectively, while $r_{x\hat{x}}$ is the sample correlation. Dividing the three parts by the total yields the bias proportion, the variance proportion, and the covariance proportion. The non-negative 'proportions' sum up to 1. According to the sources mentioned above, the *bias proportion* tells how far the mean of the forecast is from the mean of the actual series. The *variance proportion* tells how far the variation of the forecast is from the variation of the actual series. Finally, the *covariance proportion* measures the remaining unsystematic forecasting errors.

The idea is that, if the forecast is 'good', the bias and variance proportions should be small so that most of the bias should be concentrated on the covariance proportions. The informative value of the decomposition has not been accepted universally, however.

7.6 Diebold-Mariano statistics

The econometricians DIEBOLD and MARIANO were interested in a situation where a 'cheap' benchmark forecast is compared to a sophisticated forecast. A forecaster may prefer the cheap forecast up to a point where the sophisticated forecast shows its relative merits 'significantly'. It is uncertain whether this situation is common in applications and empirical projects. Usually, various forecasting methods are considered and an optimum method is then selected, while the cost of a forecast method plays little role. A complicated forecasting method may even be selected if it only achieves a slight improvement on average. Some economic projects may even show an instinctive bias toward more sophisticated and costly methods, as these demonstrate the forecasting team's skills.

CHATFIELD ranks among the few statisticians who criticized the null hypothesis of these tests *expressis verbis*. That null hypothesis would be that the expected difference in squared error (or some other loss moment) is zero. It is doubtful whether this null hypothesis is of central interest to the typical forecaster.

DIEBOLD&MARIANO assume that the precision is basically measured by $\mathrm{E}g\left(x_t - \hat{x}_{t-1}\left(1\right)\right)$ and $\mathrm{E}g\left(x_t - \tilde{x}_{t-1}\left(1\right)\right)$ for two different forecasts $\tilde{x}$ and $\hat{x}$ and a loss function $g\left(.\right)$. Under the hypothesis that the difference is zero, it can be shown that the test statistic

$$S_1 = \frac{\bar{d}}{\sqrt{m^{-1} 2\pi \hat{f}_d\left(0\right)}}$$

converges to a standard normal distribution as $m \to \infty$. Here, $\bar{d}$ denotes the sample average of $d_t = g\left(x_t - \hat{x}_{t-1}\left(1\right)\right) - g\left(x_t - \tilde{x}_{t-1}\left(1\right)\right)$. The element $\hat{f}_d\left(0\right)$ is a scale factor defined as the spectral density estimate of d_t at the frequency 0. An operable definition for $\hat{f}_d\left(0\right)$ is

$$\hat{f}_d\left(0\right) = \left(2\pi\right)^{-1} \sum_{k=-m+1}^{m-1} w\left(k, m\right) \hat{\gamma}_d\left(k\right),$$

where $\hat{\gamma}_d\left(k\right)$ is the sample autocovariance at lag k and $w\left(.\right)$ is a kernel weight function that obeys certain consistency conditions, such as $w\left(k, m\right) \to 1$ for fixed k and $m \to \infty$.

7.7 Qualitative measures

Sometimes, macro-economists maintain that they are less interested in the accuracy of a real growth forecast than in forecasting 'turning points' of the

business cycle. It is uncertain whether such turning points exist outside of the official NBER chronology and economics textbooks. Similarly, a stock market analyst may be more interested in whether a specific security price is going to rise or to fall in the immediate future than in the numerical accuracy of the price prediction for the next day. Again, it is often uncertain whether the implied picture of longer sinusoidal swings in the security price with local maxima (peaks) and local minima (troughs) corresponds to reality.

It is difficult to construct an accuracy measure for this type of loss function, at least as long as the variable of concern x is quantitative. Some economists tend to 'code' and discretize some real-valued variables, such that 'x increases', 'x decreases', and 'x remains approximately constant' become the three events or 'states'. Then, one may count the occurrences of successes and failures. The 'better' procedure is the one that yields more successes or a larger success ratio. While the winning procedure may miss the exact value of x by far, its forecast for the 'sign' of x is reliable. Here, an inherent difficulty is the definition of 'approximate constancy'. Instead of success ratios, one may also summarize this type of evidence in coincidence matrices.

A different kind of success ratio is suggested by CHATFIELD who, in comparing two methods A and B, counts the cases when A is closer to the true value and when B is closer. A forecaster who predicts many values closely and misses a few ones by far, would attain a good 'Percent Better' ratio. In this sense, CHATFIELD's 'Percent Better' is a robust criterion and comparable to the MAE. Note, however, that it is scale-independent, unlike the MAE.

7.8 Forecast bias or mean error

The sample average of the prediction errors

$$m^{-1} \sum_{t=N-m+1}^{N} \left(x_t - \hat{x}_{t-1}\left(1\right)\right)$$

is also often reported for prediction experiments. It is not a real measure of accuracy, although it contains some important information. A forecast with a large and systematic forecast bias could be improved by some straightforward adjustment. In this sense, systematic over-prediction or under-prediction points to an inefficiency, or otherwise to an asymmetric loss function.

References

[1] Anderson, T.W. (1951) 'Estimating linear restrictions on regression coefficients for multivariate normal distributions'. *Journal of the American Statistical Association* **85**, 813–823.

[2] Box, G.E.P., and G.M. Jenkins (1970) *Time Series Analysis, Forecasting, and Control*. Holden-Day.

[3] Brockwell, P.J., and R.A. Davis (1991) *Time Series: Theory and Methods*. 2nd edition, Springer-Verlag.

[4] Cartwright, N. (1995) 'Probabilities and experiments'. *Journal of Econometrics* **67**, 47–59.

[5] Chatfield, C. (2001) *Time-series Forecasting*. Chapman & Hall.

[6] Christoffersen, P.F., and Diebold, F.X. (1998) 'Cointegration and long-horizon forecasting', *Journal of Business & Economics Statistics* **16**, 450–458.

[7] Clements, M. (2005) *Evaluating Econometric Forecasts of Economic and Financial Variables*. Palgrave-Macmillan.

[8] Clements, M., and Hendry, D.F. (1998) *Forecasting economic time series*. Cambridge University Press.

[9] Clements, M., and Hendry, D.F. (1999) *Forecasting Non-Stationary Economic Time Series*. MIT Press.

[10] Dickey, D.A., and Fuller, W.A. (1979) 'Distribution of the estimators for autoregressive time series with a unit root' *Journal of the American Statistical Association* **74**, 427–431.

[11] Diebold, F.X., and Mariano, R.S. (1995) 'Comparing Predictive Accuracy' *Journal of Business and Economic Statistics* **13**, 253–263.

[12] Engle, R.F. (1992) 'Autoregressive conditional heteroskedasticity with estimates of variance of United Kingdom inflation' *Econometrica* **50**, 987–1007.

[13] Engle, R.F., Hendry, D.F., and Richard, J.-F. (1983) 'Exogeneity'. *Econometrica* **51**, 277–304.

[14] Engle, R.F., and Yoo, B.S. (1987) 'Forecasting and Testing in Cointegrated Systems', *Journal of Econometrics* **35**, 143–159.

[15] Fair, R.C. (2004) *Estimating How the Macroeconomy Works*. Harvard University Press.

[16] Fan, J., and Yao, Q. (2005) *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer.

[17] Franses, P.H. (1996) *Periodicity and Stochastic Trends in Economic Time Series*. Oxford University Press.

[18] Franses, P.H., and VanDijk, D. (2000) *Non-linear time series models in empirical finance*. Cambridge University Press.

[19] Gardner, E.S. Jr. (1985) 'Exponential Smoothing: The State of the Art' *Journal of Forecasting* **4**, 1–28.

[20] Gardner, E.S.Jr., and McKenzie, E. (1985) 'Forecasting trends in time series' *Management Science* **31**, 1237–1246.

[21] Granger, C.W.J. (1969) 'Investigating Causal Relations by Econometric Models and Cross-Spectral Methods' *Econometrica* **37**, 424–438.

[22] Granger, C.W.J. (1989) *Forecasting in Business and Economics*. Academic Press.

[23] Harvey, A.C. (1989) *Forecasting, Structural Time Series, and the Kalman Filter*. Cambridge University Press.

[24] Hylleberg, S., Engle, R.F., Granger, C.W.J., and Yoo, B.S. (1990) 'Seasonal integration and cointegration' *Journal of Econometrics* **44**, 215–238.

[25] Johansen, S. (1995) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

[26] Makridakis, S., Wheelwright, S.C., and Hyndman, R.J. (1998) *Forecasting: Methods and Applications*. 3rd edition, Wiley.

[27] Pindyck, R.S. and Rubinfeld, D.L. (1991). *Econometric Models and Economic Forecasts*. 3rd edition, McGraw-Hill.

[28] Sims, C.A. (1980) 'Macroeconomics and reality'. *Econometrics* **48**, 1–48.

[29] Taylor, S. (1986) *Modelling Financial Time Series*. John Wiley & Sons.

[30] Teräsvirta, T., Tjostheim, T., and Granger, C.W.J. *Modelling Nonlinear Economic Time Series*. Oxford.

[31] Tong, H. (1990) *Non-linear Time Series: A Dynamical Systems Approach*. Oxford University Press.