Econometric Forecasting

Robert M. Kunst
robert.kunst@univie.ac.at

University of Vienna
and
Institute for Advanced Studies Vienna

November 10, 2012

Outline

The basic problem

Given some data (observations) on a (possibly multivariate) variable x, i.e. $x_1, \ldots, x_N$, we want to find a good approximation to the (as yet 'unknown') observation x_{N+h}. We use CHATFIELD's notation: $\hat{x}_N(h)$ is a h–step forecast for x_{N+h} given observations (a time series) until and including x_N.

The information set available at $t = N$ for the forecast necessarily includes the observed time series but it may be much larger in practice.

Forecasting and predicting

To many authors, *forecasting* and *prediction* are equivalent. Some authors distinguish the terms: prediction is the technical word, forecasting relates predictions to the substance-matter environment.

Clements and Hendry define: *predictability* is a theoretical property—unconditional and conditional distributions differ—, *forecastability* is the possibility that this property can be exploited in practice.

The words *prognosis* and *projection* are related but their usage is more restricted.

The participle *forecasted* is incorrect but ubiquitous in the literature.

Aims of forecasting

1. Curiosity: we want to know about the future;
2. Decision making: may specify a loss function for prediction, but fully developed examples are rare;
3. Policy evaluation: modified policy may change expectations and thus model behavior (Lucas critique);
4. Model evaluation: quality of models as descriptive tools and of model-based predictions may differ.

Forecasts: classification according to objectivity

1. *magical*: no recognizable cause-effect relationship (oracles);
2. *subjective*: judgmental forecasts, Delphi method (averaging over questionnaires);
3. *objective*: univariate (one time series only), multivariate (several variables summarized in a vector).

In practice, most institutional forecasts use a mix of subjective and objective elements.

Forecasts: classification according to being automatic

1. *automatic* methods: forecasts based on a well-defined rule without any further user intervention;
2. *non-automatic* methods: forecasts require some action (decision) by the user (e.g. FAIR's 'add factors').

Most procedures studied in the literature are automatic. In practice, forecasts almost always utilize some user intervention.

Model-free and model-based prediction

1. *model-free procedures*: extrapolation by free hand, exponential smoothing, trend fitting;
2. *model-based procedures*: data-driven (time series) or theory-driven (e.g. econometric models).

Some extrapolation methods can be justified by time-series models. It is easier to evaluate model-based procedures, as the models can be simulated. With actual data, extrapolation can be a surprisingly good benchmark.

Model-based forecasting: classification by complexity

A Nature of the DGP
- i Stationary DGP
- ii Co-integrated DGP
- iii Evolutionary, non-stationary DGP

B Knowledge level
- i Known DGP; known θ
- ii Known DGP; unknown θ
- iii Unknown DGP; unknown θ

C Dimensionality of the system
- i Scalar process
- ii Closed vector process
- iii Open vector process

D Form of analysis
- i Asymptotic analysis
- ii Finite-sample analysis

E Forecast horizon
- i 1–step
- ii Multi-step

F Linearity of the system
- i Linear
- ii Non-linear

Self-fulfilling and self-defeating forecasts

If decisions are based on forecasts, forecasts may affect the forecasted variables. Effects can be positive or negative:

1. *self-fulfilling* forecasts: a bad growth forecast may cause pessimism and decrease demand; a high inflation forecast may raise incentives for wage bargaining;
2. *self-defeating* forecasts: a high unemployment forecast may cause active labor market policies; a high inflation forecast may cause central banks to implement anti-inflationary policies.

A good excuse for inaccurate economic forecasts?

Out-of-sample and in-sample forecasts

1. a true *out-of-sample* forecast $\hat{x}_N(h)$ only uses information over the time range $t \leq N$. If it is model-based, all parameters are estimated for $t \leq N$, and this includes data-based elements of model specification;
2. an *in-sample* forecast uses information over $t \leq N + h$. Such information may be exogenous variables, or a model is fitted to a time range ending even after $N + h$. Forecast errors will be *residuals*, not true prediction errors.

In forecasting, good performance in out-of-sample prediction is viewed as the acid test for a good forecast model.

Forecast failure

- Demographic projections are relatively accurate, even for longer terms;
- the accuracy of meteorological forecasts has improved over the last decades;
- demand forecasts for new products have often failed completely (computers, TV sets), and they will continue to do so;
- speculative markets are very difficult to forecast (Sir Clive Granger);
- short-run macroeconomic forecasts **did not improve** over the last decades (excuses: changing environment, human action, . . .).

Literature on forecasting

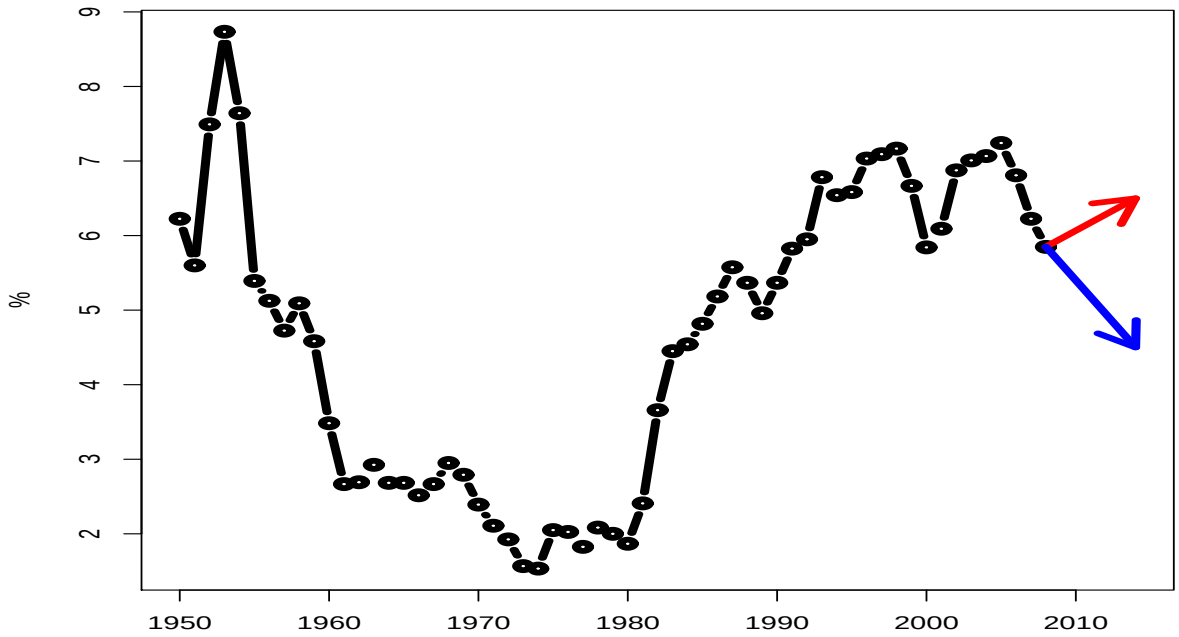
- Chatfield, C. (2001) *Time Series Forecasting*, Chapman& Hall: accessible survey
- Makridakis, S., Wheelwright, S.C., and R.J. Hyndman (1998) *Forecasting: Methods and Applications*, Wiley: introduction for business economists
- Clements, M., and D.F. Hendry (1998) *Forecasting economic time series*, Cambridge University Press: academic
- Clements, M., and D.F. Hendry (1999) *Forecasting Non-Stationary Economic Time Series*, MIT Press: a second part
- Clements, M. (2005) *Evaluating Econometric Forecasts of Economic and Financial Variables*, Palgrave-Macmillan: a summary
- Journals *Journal of Forecasting* and *International Journal of Forecasting*: the state of the art

The free-hand method

Instinctively, we attempt to fit smoothed curves through time series and to extrapolate them at the end to generate a 'forecast'.

The forecast will be very subjective and depends on the amount of smoothing.

Example: the Austrian unemployment rate



Extrapolation of the last few years yields a low prediction, smoothing over the last decade yields an upward direction.

Exponential smoothing: the idea

The variants of *exponential smoothing* are sophisticated objective versions of the free-hand method.

Technical terminology:

- a causal *filter* determines the (filtered) value of a data point $\hat{x}_N$ from observations x_t, $t < N$;
- a *smoother* determines the (smoothed) value of a data point $\hat{x}_N$ from observations x_t, $t < N$ and $t \geq N$.

Extrapolation calculates a smoother in order to apply a causal filter afterwards.

Single exponential smoothing (SES)

SES determines the filtered value $\hat{x}_t$ from a weighted average over a past observation and a past filtered value:

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

The constant $\alpha \in (0, 1)$ is a *damping* or *smoothing factor*. This equation is called the 'recurrence form' of SES. Note that a smoothed past needs to be known here.

Properties of SES: sum representation

Repeated substitution yields

$$\hat{x}_t = \alpha \sum_{j=0}^{t-1} (1-\alpha)^j x_{t-j} + (1-\alpha)^t \hat{x}_0,$$

often given without the last term, assuming $\hat{x}_0 = 0$. Large α implies strong 'discounting' of the past and weak smoothing. In any case, the past enters with geometrically declining weights.

Properties of SES: error-correction form

Re-arranging the recurrence form yields the 'error-correction form' of SES:

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

'The new smoothed value is the old smoothed value plus some correction for the prediction error', if we interpret $\hat{x}_{t-1}$ as a forecast for x_t.

How to choose α

Two main suggestions:

1. Choose α from the interval $[0.1, 0.3]$. 0.1 implies strong smoothing, 0.3 implies weak smoothing.
2. Determine α by optimizing prediction over the sample, i.e.

$$\min_{\alpha} \Sigma \left(x_{t+1} - \hat{x}_t\right)^2$$

Problem with second option: α cannot be an estimate, as no generating model is assumed.

How to choose a starting $\hat{x}$

Three options:

1. Start from the actual data $\hat{x}_1 = x_1$. Problems for small α and small samples;
2. Start from a sample average $\hat{x}_1 = \bar{x}$. Variants calculate means from parts of the sample;
3. *backcasting*: run the SES filter backward to determine $\hat{x}_{t-1}$ from x_t and $\hat{x}_t$, using a starting value later in the series, finally determine $\hat{x}_1$.

SES and discounted least squares

If the level of a series changes slowly, it may be advantageous to determine a location μ not by least squares

$$\min_{\mu} \sum_{j=0}^{t-1} (x_t - \mu)^2,$$

but by *discounted least squares*

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

for a 'local' mean in t. This yields SES for $\beta = 1 - \alpha$.

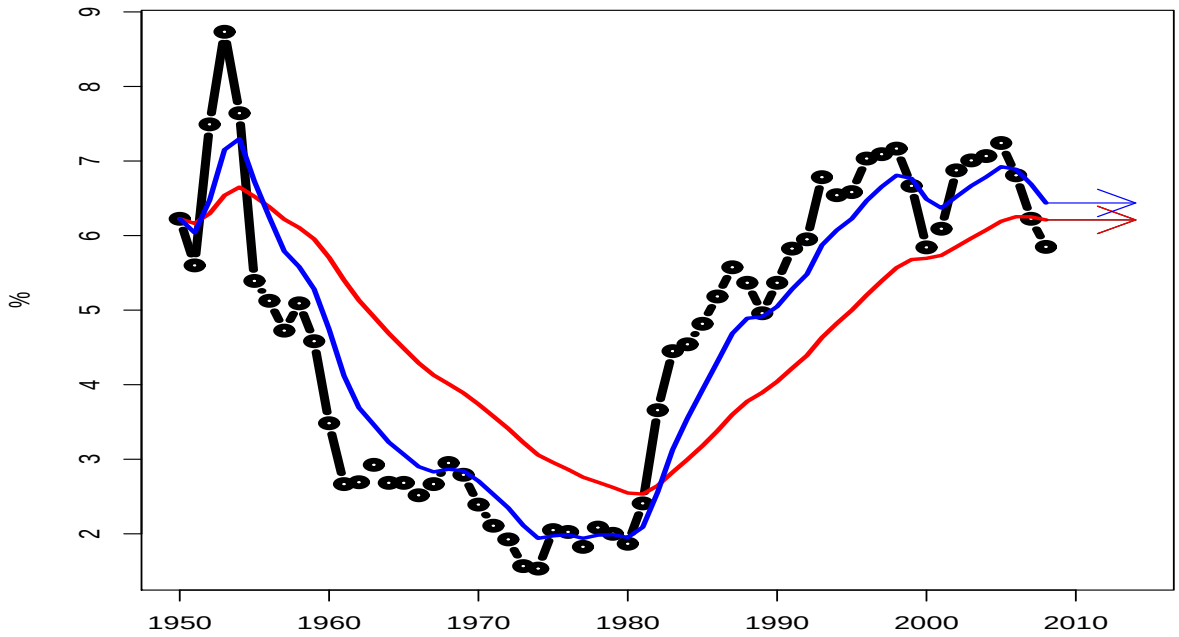
The SES forecast

SES defines a forecast $\hat{x}_N(h)$ by flat extrapolation:

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

This is not appropriate for trending variables.

Example: the Austrian unemployment rate



SES for the unemployment rate (red $\alpha = 0.1$, blue $\alpha = 0.3$) and implied forecasts.

Double exponential smoothing (DES)

Double exponential smoothing runs SES twice:

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

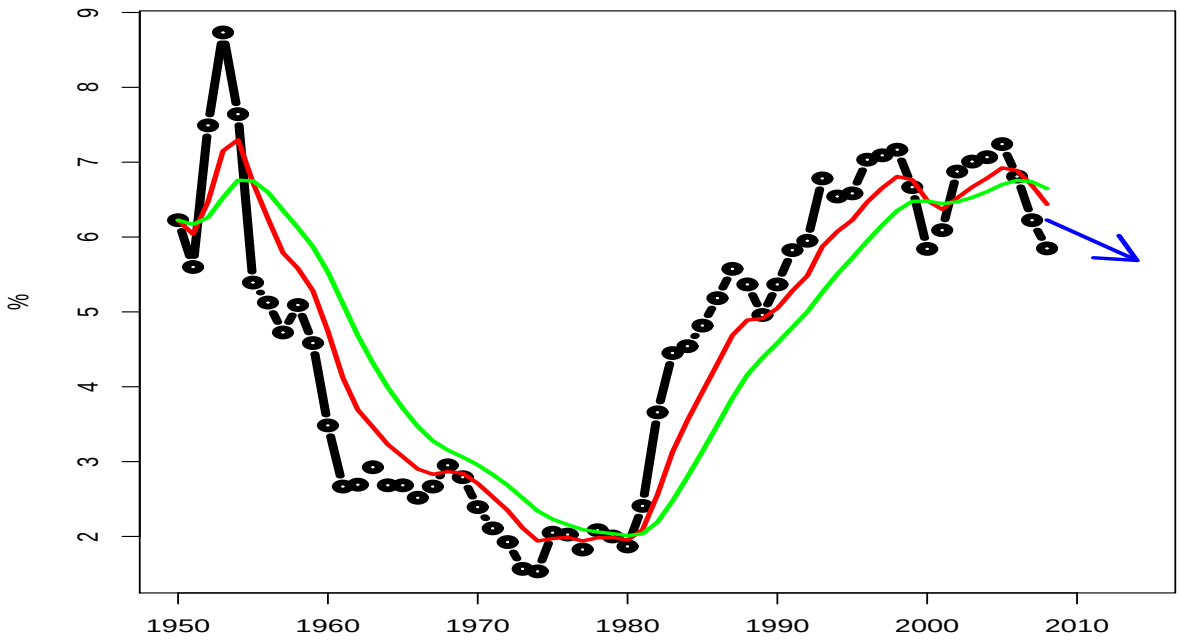
with L a *local level* and T a *local trend* estimate. Both L and T are smoothed versions of the data.

The DES forecast

h–step forecasts extrapolate the last observations on L_t and T_t:

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

Example: DES on the Austrian unemployment rate



DES for the unemployment rate ($\alpha = 0.3$, red L, green T) and implied forecasts (blue).

Tentative assessment of SES and DES

- SES and DES are quick and simple procedures that often perform well;
- SES can be shown to be equivalent to a time-series forecast based on an ARIMA(0,1,1) model, i.e. MA on first differences;
- DES is equivalent to a time-series forecast for a specific ARIMA(0,2,2) model, i.e. MA(2) on second differences with parameter restrictions on MA coefficients: not a very plausible model;
- For this reason, many forecasters avoid DES and use the more flexible Holt-Winters methods instead.

Holt's linear trend method

Holt's method generalizes DES and introduces a second tuning parameter. It has two recursion equations, for local trend (or rather 'slope') T and local level L:

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

L averages data and 'forecast', T averages old slope and new slope estimate from L. Meaning of T differs from DES! A very popular method.

Forecasting using Holt's method

The standard definition for h–step forecasts is

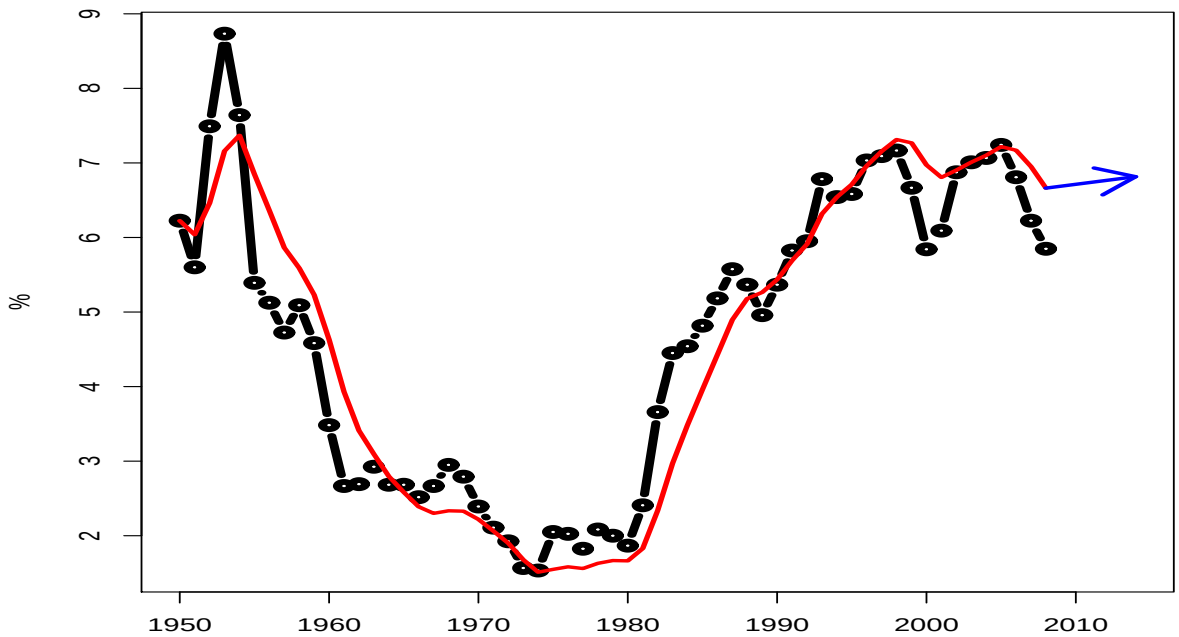
$$\hat{x}_N(h) = L_N + hT_N,$$

such that $\hat{x}_{t-1}(1) = L_{t-1} + T_{t-1}$ is a smoothed version of x_t. Gardner& McKenzie suggest to forecast from Holt's method via

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

This forecast corresponds to an ARIMA(1,1,2) generating model, while Holt's method relies on ARIMA(0,2,2). Chatfield warns that all smoothing extrapolations are not genuinely justified by prediction in time-series models. They would imply absurd parameter restrictions.

Example: Holt on the Austrian unemployment rate



Holt procedure applied to the unemployment rate ($\alpha = 0.3$, $\gamma = 0.1$, red L) and implied forecasts (blue). Trend changes slowly and still points upward at the end. Least-squares fitting would suggest even more extreme parameter values.

Holt-Winters seasonal method

Quarterly and monthly economic data often has considerable seasonal variation. Traditional seasonal models distinguish multiplicative seasonality (seasonal factors) and additive seasonality (seasonal dummy intercepts). The *Holt-Winters* method allows for slow changes in these seasonal factors and intercepts.

Holt-Winters: the recursions

Multiplicative version:

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

Additive version:

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

Remarks on Holt-Winters

- Typically, $s = 4$ or $s = 12$.
- The procedure needs three smoothing parameters that are often determined by least-squares fitting. Low γ prescribes a time-constant, deterministic seasonal cycle.
- Convenient starting values for T are sample averages over Δx. For S, one may use s averages over the specific season. Averages may be restricted to a first portion of the sample.
- While the Holt method corresponds to an ARIMA(0,2,2) generating model, there is no simple time-series model that justifies Holt-Winters. Nonetheless, the method works well in practice.

The Holt-Winters prediction formulae

The last available seasonal pattern is extrapolated into the future.

Multiplicative version:

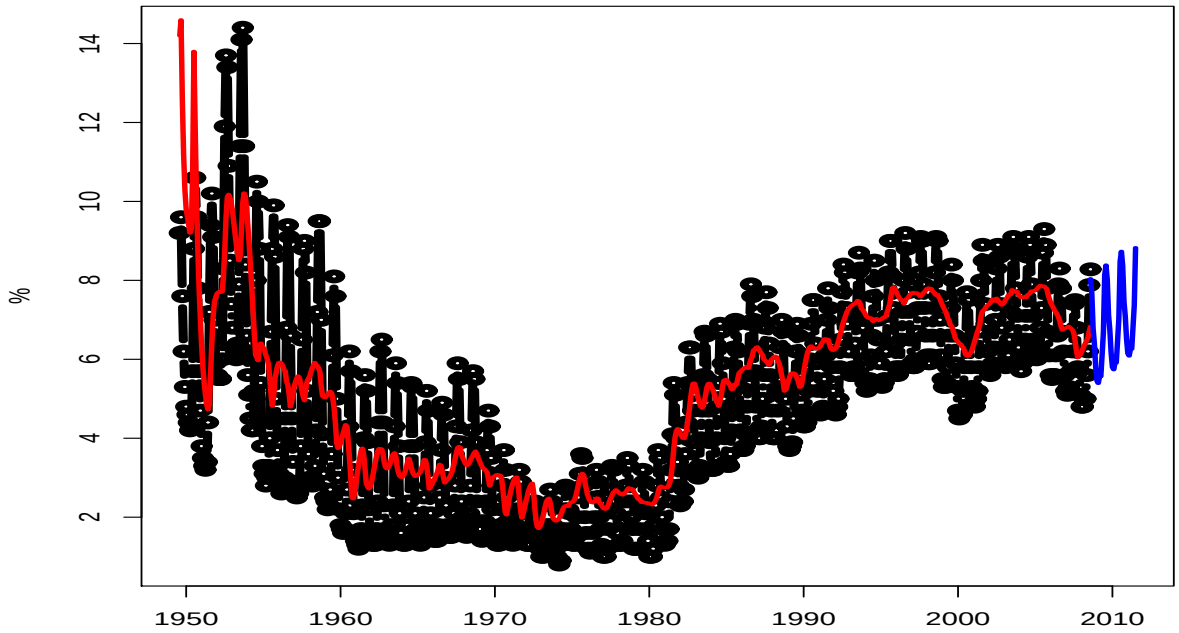
$$\hat{x}_N(k) = (L_N + T_N k)\, S_{N+k-s},$$

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

Additive version:

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

Example: Holt-Winters on the Austrian unemployment rate



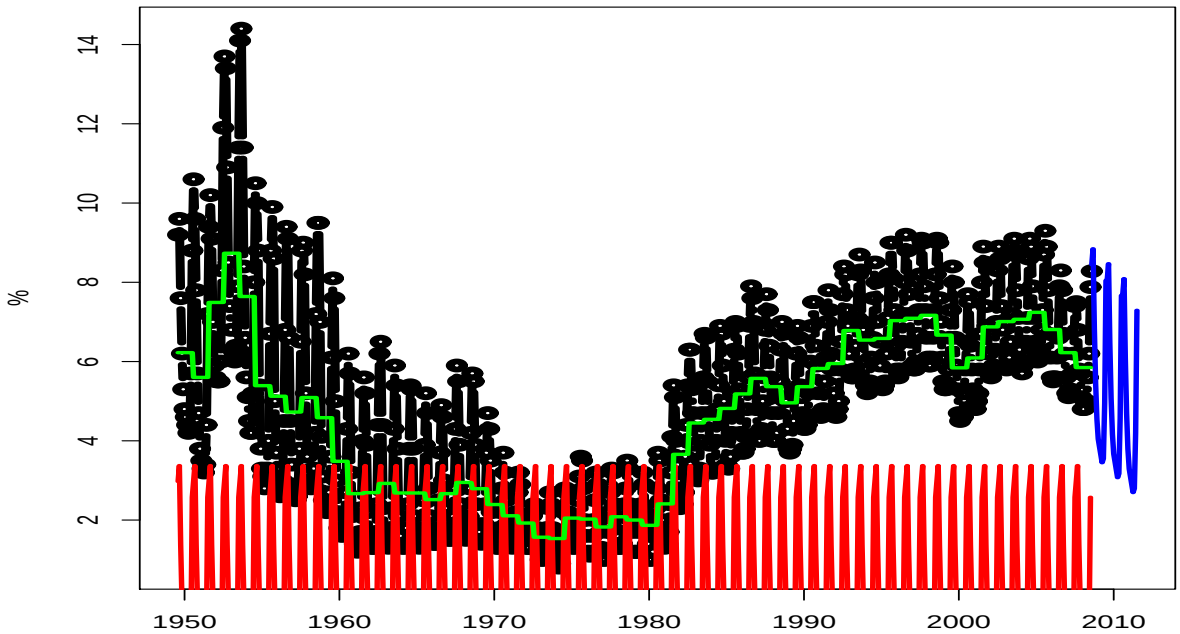
Additive Holt-Winters ($\alpha = 0.3$, $\beta = 0.1$, $\gamma = 0.5$) applied to the monthly Austrian unemployment series. Red: shifted L, blue: forecast.

The Brockwell & Davis small-trends method

The time-series researchers Brockwell & Davis suggested an appealingly simple alternative to the complex Holt-Winters algorithm:

1. Calculate annual averages for the series, interpret them as 'trend', and subtract the trend from the observed x_t to yield a seasonal, but not trending $\tilde{x}_t$;
2. Calculate averages for each season in $\tilde{x}_t$ over all years, which yields an estimate of the seasonal cycle;
3. Extrapolate the trend (which one?) plus cycle into the future to obtain a forecast.

Example: Brockwell-Davis small-trends on Austrian unemployment rate



Brockwell-Davis small-trends method applied to the monthly Austrian unemployment series. Red: seasonal cycle, green: 'trend', blue: forecast.

Remarks on the small-trends method

- The method assumes a time-constant seasonal cycle, which often does not appear to be realistic;
- The method fails if trends really play a role. One may modify the method fitting simple functions of time and extrapolating them.

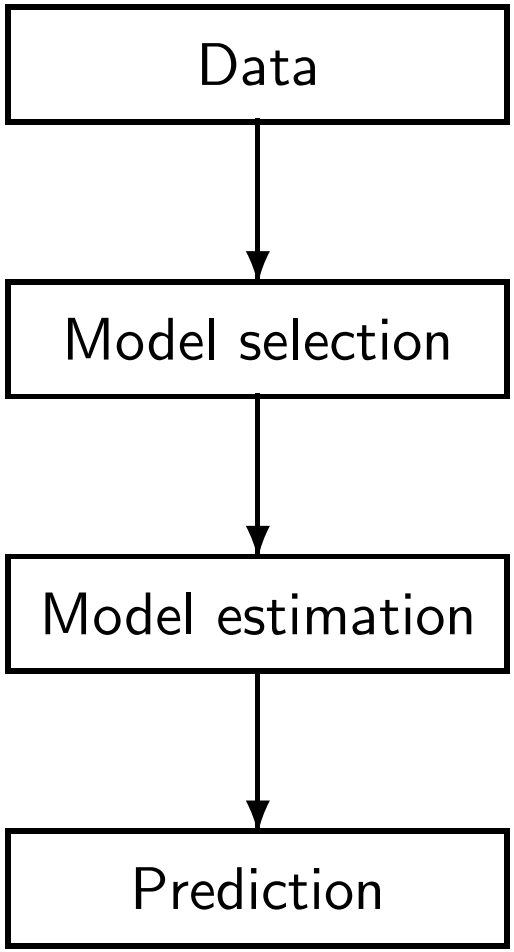
Forecasting using time-series models

These methods tentatively assume a data-generating process, the degree of belief in their model varies among researchers: "All models are wrong, some are useful" [G.E.P. Box]

Notes:

- 'useful' may refer to forecasting;
- 'Model' usually refers to a parametric model class. The best-fitting or true parameter value is unknown and has to be estimated;
- knowing the true model class does not guarantee the best forecasting performance if parameters have to be estimated. The wrong model class may outperform the wrong parameter in the true class: the true model class can be 'useless';
- simple linear time-series models are good forecasters even if the data-generation process is nonlinear.

The stages of time-series prediction



The general form of time-series models

The current X_t depends on its past and on an error:

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

where g is a nonlinear or linear function, θ is an unknown parameter, (ε_t) is an unobserved error process.
(ε_t) is often assumed *i.i.d.* but is at least a *martingale-difference sequence* (MDS) defined by

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

$\mathfrak{I}_{t-1}$ is an information set containing the process past. White noise (uncorrelated) (ε_t) is not sufficient for prediction!

Prediction using a time-series model

Suppose θ is known. Then

$$\begin{aligned} \mathrm{E}(X_t|\mathfrak{I}_{t-1}) &= g\left(X_{t-1}, X_{t-2}, \ldots; \theta\right) + \mathrm{E}(\varepsilon_t|\mathfrak{I}_{t-1}) \\ &= g\left(X_{t-1}, X_{t-2}, \ldots; \theta\right) \end{aligned}$$

is a convenient forecast $\hat{X}_{t-1}(1)$. It is easily shown that it minimizes the expected squared prediction error $\mathrm{E}(e_t)^2$ with $e_t = X_t - \hat{X}_t$ among all feasible $\hat{X}_t$.

If θ is unknown, it is estimated from the sample and plugged in, as if it were known. If the model class is correct and the sample is large, many authors claim that the reduction in accuracy is minor.

General nonlinear time-series models

Many nonlinear models do not obey the linear-errors scheme. The general form is:

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

Here, even if θ were known, we have

$$\mathrm{E}(X_t|\mathfrak{I}_{t-1}) \neq g\left(X_{t-1}, X_{t-2}, \ldots; 0; \theta\right).$$

The only correct solution is *stochastic prediction*.

Stochastic prediction

1. Choose a generating distribution for the errors ε_t;
2. Draw from a random processor and generate J replications of

$$\tilde{X}_t^{(j)} = g\left(X_{t-1}, X_{t-2}, \ldots; \varepsilon^{(j)}; \hat{\theta}\right);$$

3. Average over the J replications

$$\hat{X}_{t-1}^{J}(1) = \frac{1}{J}\sum_{j=1}^{J} \tilde{X}_t^{(j)},$$

   to approximate the expectation (law of large numbers, LLN);
4. Analogous steps can be taken for h–step prediction with $h > 1$.

From what distribution to draw

Two main suggestions:

1. Fit a parametric model, for example normal distribution, to the residuals and estimate the parameters: parametric bootstrapping, Monte Carlo;
2. Draw from a discrete uniform law over the sample residuals: (nonparametric) bootstrapping.

Linear models with rational lag functions

ARMA (*autoregressive moving average*) models of the form

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

are the most popular time-series models for data-based prediction. Their popularity may still be due to the book by Box and Jenkins (1970,1976). We repeat the main results for special cases (AR, MA) in brief.

The autoregressive model

The AR(p) model

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

can be estimated by least squares. Note that ε_t is specified as MDS, not simply as white noise, in forecasting applications.

Stability of the AR model

The AR(p) model is said to be *stable* if it permits a stationary process that satisfies the equation and if future depends on the past in that solution. For example, $X_t = 2X_{t-1} + \varepsilon_t$ has a stationary solution that is useless for forecasting. A stable model is also called *asymptotically stationary*.

The AR(p) model is stable if its characteristic polynomial equation

$$1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = 0$$

has only roots *greater* than one in modulus. For small p, this property is easily checked by hand.

Determining the lag order p of an AR model

Time-series analysis knows three approaches for lag-order search:

1. Plot the empirical *partial autocorrelation function* (PACF) $\rho_P(k)$ that should differ from 0 for $k \leq p$ and equal 0 for $k > p$ (recommended by Box& Jenkins);
2. fit AR(p) models for different p and test residuals for white noise: choose the smallest p such that the test is passed (unreliable);
3. fit AR(p) models for different p and calculate information criteria $IC(p)$ for each model: choose the p that minimizes the criterion.

The information-criterion approach is the most suitable one for forecasting.

Information criteria

There are two main classes of information criteria:

1. *Consistent* criteria: as the sample size $N \to \infty$, the true lag order tends to be found with probability one—for example, Schwarz' BIC;
2. *efficient* criteria: as the sample size $N \to \infty$, the forecast based on the selected model minimizes the expected mean-squared error—for example, Akaike's AIC

$$AIC = \log \hat{\sigma}^2(p) + \frac{2p}{N},$$

with $\hat{\sigma}^2(p)$ an estimated errors variance from an AR(p) model.

If the aim is forecasting, criteria of the second class, which includes AIC, AIC_u, AIC_c, FPE, may be a natural choice.

Forecasting from an AR(p) model

Suppose the AR(p) model has generated the data and the parameters are known. Then:

$$\mathrm{E}(X_{t+1}|\mathfrak{I}_t) = \phi_1 X_t + \phi_2 X_{t-1} + \ldots + \phi_p X_{t-p+1} + \mathrm{E}(\varepsilon_{t+1}|\mathfrak{I}_t),$$

and hence, as ε is MDS,

$$\hat{X}_t(1) = \phi_1 X_t + \phi_2 X_{t-1} + \ldots + \phi_p X_{t-p+1}.$$

If parameters are unknown and also p has been determined empirically, plugging in $\hat{\phi}_j$ yields a feasible approximation that is acceptable for larger N.

Multi-step forecasting from an AR(p) model

Two-step forecasts are obtained by iteration, plugging in one-step forecasts:

$$\mathrm{E}(X_{t+2}|\mathfrak{I}_t) = \phi_1 \mathrm{E}(X_{t+1}|\mathfrak{I}_t) + \phi_2 X_t + \ldots + \phi_p X_{t-p+2} + \mathrm{E}(\varepsilon_{t+2}|\mathfrak{I}_t),$$

or

$$\hat{X}_t(2) = \phi_1 \hat{X}_t(1) + \phi_2 X_t + \ldots + \phi_p X_{t-p+2},$$

and this iteration can be continued for larger horizons.

Multi-step forecasting: direct modeling

In iterated plugging-in, the forecast $\hat{X}_t(h)$ is formed as

$$\hat{X}_t(h) = \zeta_h X_t + \zeta_{h+1} X_{t-1} + \ldots + \zeta_{h+p-1} X_{t-p+1},$$

with ζ_j depending on $\phi_1, \ldots, \phi_p$. Alternatively, one may fit models of the type

$$X_t = \phi_h X_{t-h} + \ldots + \phi_p X_{t-p} + \varepsilon_t$$

to the sample and use

$$\hat{X}_t(h) = \phi_h X_t + \ldots + \phi_p X_{t-p+h}.$$

The relative merits of this *direct modeling* method are an issue of ongoing research. For small h and correctly specified models, iterated forecasting can be shown to be better.

The moving-average model

The MA(q) model

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

is to be estimated by a non-linear least-squares procedure. Again note that ε_t is specified as MDS in forecasting applications.

Stability of the MA model

The MA(q) model is always stable. Excluding some q starting values, MA processes are stationary, not only asymptotically stationary.

Evaluation of the characteristic polynomial equation

$$1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q = 0$$

is nevertheless helpful. If it has only roots *greater* than one in modulus, there exists a convergent infinite-order autoregressive representation

$$\sum_{j=0}^{\infty} \psi_j X_{t-j} = \varepsilon_t,$$

which can be useful for prediction. In this case, the MA(q) model is said to be *invertible*.

What does non-invertibility imply?

Two cases:

1. If any of the polynomial roots have modulus *exactly equal one*, prediction becomes very difficult. Sometimes, this non-invertibility is due to pre-processing the data by filtering, seasonal adjustment, differencing;
2. If any roots have modulus *less than one*, there exists an observationally equivalent MA model with all roots larger than one. This non-invertibility is due to a non-optimal estimation routine.

Determining the lag order q of an MA model

Again, time-series analysis knows three approaches for lag-order search:

1. Plot the empirical *autocorrelation function* (ACF) or *correlogram* $\rho(k)$ that should differ from 0 for $k \leq q$ and equal 0 for $k > q$ (recommended by Box& Jenkins);
2. fit MA(q) models for different q and test residuals for white noise: choose the smallest q such that the test is passed (unreliable);
3. fit MA(q) models for different q and calculate information criteria $IC(q)$ for each model: choose the q that minimizes the criterion.

Again, the IC approach is the most suitable one for forecasting, and there may be a preference for using 'efficient' criteria, such as AIC.

Forecasting from an MA(q) model

Suppose the MA(q) model has generated the data and the parameters are known. Then, one may reconstruct true ε_t and use:

$$\mathrm{E}(X_{t+1}|\mathfrak{I}_t) = \theta_1\varepsilon_t + \theta_2\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q+1} + \mathrm{E}(\varepsilon_{t+1}|\mathfrak{I}_t),$$

and hence, as ε is MDS,

$$\hat{X}_t(1) = \theta_1\varepsilon_t + \theta_2\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q+1}.$$

If parameters are unknown, q is determined empirically, and ε_t must be estimated, one may still plug in estimates. Alternatively, program routines may use the 'inverted' AR(∞) model and cut off the sum at some large value.

Multi-step forecasting from an MA(q) model

Two-step forecasts are simple in principle, according to:

$$\mathrm{E}(X_{t+2}|\mathfrak{I}_t) = \theta_1 \mathrm{E}(\varepsilon_{t+1}|\mathfrak{I}_t) + \theta_2 \varepsilon_t + \ldots + \theta_q \varepsilon_{t-q+2} + \mathrm{E}(\varepsilon_{t+2}|\mathfrak{I}_t),$$

or

$$\hat{X}_t(2) = \theta_2 \varepsilon_t + \ldots + \theta_q \varepsilon_{t-q+2},$$

and this scheme can be continued for larger horizons h. For $h > q$, $\hat{X}_t(h) = 0$. MA processes are *finite dependent*.

The autoregressive moving-average model

The general ARMA(p, q) model

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

is to be estimated by a non-linear least-squares procedure. Again note that ε_t is specified as MDS in forecasting applications.

Stability of the ARMA model

The ARMA(p, q) model inherits its properties from its AR and MA components.

1. For a unique definition, the characteristic polynomials for the AR and MA parts must not have common zeros, otherwise a simpler representation ARMA($p-1, q-1$) exists and the additional parameters cannot be estimated;
2. under condition # 1, if the AR polynomial has only roots larger than one, the ARMA model is stable;
3. under condition # 1, if the MA polynomial has only roots larger than one, the ARMA model is invertible, i.e. there is an AR(∞) representation.

Determining the lag orders p and q of an ARMA model

In principle, time-series analysis knows three approaches for lag-order search:

1. Plot advanced tools such as the empirical *extended autocorrelation function* (EACF) and guess a good combination of lag orders by visual inspection (rarely used);
2. fit ARMA(p, q) models for different p and q and test residuals for white noise: choose the smallest p and q such that the test is passed (unreliable);
3. fit ARMA(p, q) models for different p and q and calculate information criteria $IC(q)$ for each model: choose the pair (p, q) that minimizes the criterion.

Again, there may be a preference for using 'efficient' criteria, such as AIC.

Forecasting from an ARMA(p, q) model

Forecasting must proceed carefully, using a variant of the method used in AR models. Suppose an ARMA$(2, 2)$ model has generated the data and the parameters are known. Then, one may reconstruct true ε_t and use:

$$\mathrm{E}(X_{t+1}|\mathfrak{I}_t) = \phi_1 X_t + \phi_2 X_{t-1} + \mathrm{E}(\varepsilon_{t+1}|\mathfrak{I}_t) + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1},$$

and hence, as ε is MDS,

$$\hat{X}_t(1) = \phi_1 X_t + \phi_2 X_{t-1} + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1}.$$

In practice, parameters are estimated, p and q are determined empirically, and ε_t must be estimated, and these estimates are plugged in. Alternatively, program routines may use the 'inverted' AR(∞) model and cut off the sum at some large value.

Multi-step forecasting from an ARMA(p, q) model

Two-step forecasts are obtained by plugging in one-step forecasts for the true values, according to:

$$\begin{aligned} \mathrm{E}(X_{t+2}|\mathfrak{I}_t) &= \phi_1 \mathrm{E}(X_{t+1}|\mathfrak{I}_t) + \phi_2 X_t + \\ &\quad \mathrm{E}(\varepsilon_{t+2}|\mathfrak{I}_t) + \theta_1 \mathrm{E}(\varepsilon_{t+1}|\mathfrak{I}_t) + \theta_2 \varepsilon_t, \end{aligned}$$

or

$$\hat{X}_t(2) = \phi_1 \hat{X}_t(1) + \phi_2 X_t + \theta_2 \varepsilon_t,$$

and this scheme can be continued for larger horizons h. For $h > q$ (here, $q = 2$), the MA part disappears.

Integrated models

The class of *integrated models* is the most popular class of *non-stationary* time-series models. An integrated process (X_t) is defined by the property that it is not stationary but d-th order differences $(\Delta^d X_t)$ are stationary. Only $d = 1$ and $d = 2$ are of empirical interest.

The class of integrated processes is a very special class of non-stationary processes. They model near-polynomial and random-walk trends well but not structural breaks, outliers, increasing variation, and other observed non-stationary features. Note that *data* cannot be *non-stationary*.

Notation

Box & Jenkins called a process X_t ARIMA(p, d, q) if $\Delta^d X_t$ is a stable and well-defined ARMA(p, q) process but $\Delta^{d-1} X_t$ is not.

Engle & Granger called a process X_t *d-th order integrated*, in symbols $\mathbf{I}(d)$ if $\Delta^d X_t$ is stationary but $\Delta^{d-1} X_t$ is not stationary. This is slightly more general.

Note $\Delta X_t = X_t - X_{t-1}$ and $\Delta^2 X_t = X_t - 2X_{t-1} + X_{t-2}$.

How to handle data from integrated models

If data stem from an $I(d)$ process, the idea is:

1. take d–th order differences;
2. fit ARMA models to the differenced data;
3. forecast according to the identified ARMA structures;
4. possibly integrate back (accumulate) to obtain forecasts for the original variable.

How to decide whether data stem from integrated processes

Two main ideas:

1. Box & Jenkins suggest to consider the correlogram. If it decays too slowly, take differences. Use the differencing order that makes the correlogram as simple as possible: *over-differencing* would make it more volatile;
2. Most economists today base this decision on the test by Dickey & Fuller and comparable tests. The null hypothesis is the 'unit root': if the test does not reject, take differences.

Is it so good to use DF tests in forecasting?

The answer is uncertain. Note that

1. Unit-root tests have *low power* and tend to support the null, i.e. differencing;
2. in finite samples, it is not certain that the statistical unit-root decision and the optimal procedure for forecasting coincide. In other words, ΔX_t may be easier to forecast even if X_t is stationary;
3. there is no general guideline for the significance level of the test that optimizes forecasting performance;
4. Clements and Hendry provide evidence that differencing may improve forecasting performance in the presence of breaks and outliers, even though unit-root tests reject.

Nonlinear forecasting models

General nonlinear models are rarely used in econometric forecasting. Three specific classes are popular:

1. ARCH models: autoregressive conditional heteroskedasticity;
2. threshold models;
3. artificial neural networks (ANN).

ARCH

The original ARCH model by Engle (1982) assumes, in its simplest form, that X_t is white noise. If X_t follows

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

(X_t) is said to be an ARCH(r) process. The model is stable with $\mathrm{var} X_t < \infty$ if

1. $\alpha_0 > 0$;
2. $\alpha_j \geq 0, 0 < j \leq r$;
3. $\sum_{j=1}^{r} \alpha_j < 1$.

GARCH

The most popular ARCH generalization today is still the GARCH model by Bollerslev. The lagged unobserved conditional variance serves to reflect an ARMA–type geometric decay of volatility shocks. The GARCH(1,1) model reads

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

which models well (log differences of) near random walks in the financial world.

ARMA-ARCH

To the time-series forecaster who models serially correlated variables, the most interesting extensions are ARMA-ARCH models with non-trivial *mean equation* and an ARCH–type *variance equation*, for example:

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

Note that this form already appears in the work of Engle (1982), where monthly U.K. inflation was modelled.

Stable ARMA-ARCH models

If the mean equation fulfills the usual ARMA stability conditions and the variance equation fulfills the ARCH stability conditions, ε_t is white noise and (X_t) is a stationary *homoskedastic* process. The models view h_t ('volatility', 'risk'?) as time-dependent but unconditional $\mathrm{var} X_t$ as time-constant.

ARCH models for forecasting: worth the additional work?

1. Forecasts for the 'level' of X_t are as good as the mean equation, the ARCH parameters only enter indirectly, they serve to estimate e.g. ϕ more efficiently and they specify the standard error of $\hat{\phi}$;
2. for any data with monthly or lower frequency, modelling the ARCH part is not worth the work, gains in efficiency are low;
3. it is tempting to forecast X_t^2 on the basis of the ARCH model but such forecasts are often surprisingly poor;
4. current research opines that a systematic forecast of 'local risk' aiming at commercial advice to risk-conscious traders based on ARCH models is not possible or at least 'very difficult' (Granger).

Threshold models: the idea

Economic agents may react differently to positive and negative shocks, to small and to large shocks.

It may make sense to consider models such as:

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

where $r_k = \infty$ and $r_0 = -\infty$. These models are called SETAR (self-exciting threshold autoregressive) models.

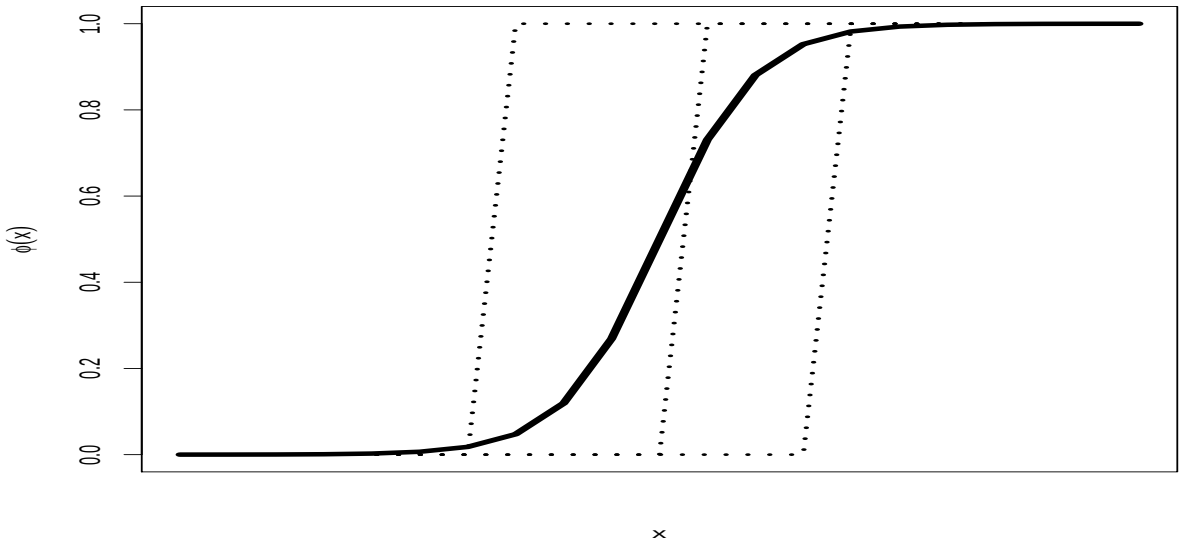
Stability of SETAR models

- First-order SETAR models are stable if the 'outer' regimes have $|\phi_j| < 1$ (sufficient only);
- higher-order SETAR models have very complex stability conditions;
- most empirical modelling is done for two or three regimes.

The benefits of SETAR for prediction

- Threshold reaction is often found in macroeconomics;
- thresholds r_j and coefficients for rarely observed regimes need extremely large samples for reliable estimation;
- forecasting must be based on stochastic prediction, as the model is nonlinear, distributions of ε_t play a key role;
- the models may forecast poorly even when they are the data-generating processes and may be outperformed by linear models.

Neural networks: the idea



Single neurons are activated according to 0-1 functions. The sum of many neurons allows a smooth transition from 'no reaction' (0) to 'unbearable pain' (1). Basing models on sigmoid functions rather than linear ones may often be more realistic.

Layers of neurons

Neurons may be linked to further neurons (synapses), which permits 'multiple layers'. The simplest version of neural nets used in practice has just one 'hidden layer' of such synapses. Assume the 'stimulus' or 'input' are past x, and the 'reaction' or 'output' is current x. This forecast net function follows CHATFIELD:

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

where all ϕ_h are sigmoid functions.

Training the net

Weights and numbers of layers are typically optimized over an estimation interval (*training set*) and are then used for prediction based on the identified *architecture*. Extending the training set to an intermediate sample to update the weights is called *learning*.

Neural nets are really just a class of nonlinear time-series models. Their reported forecasting successes may be rooted in the usage of sigmoid reaction functions.

State-space modelling: the idea

Assume an unobserved multivariate state θ_t determines the observed variable of interest X_t. There is an *observation equation*

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

with a noise term n_t and a vector of linear weights h_t. The state behaves according to a *transition equation*

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

with square matrix G and another noise term w_t.

In this general form, the model cannot be identified.

Example: autoregressive models in state-space form

For example, assume $p = 4$. Define

$$G_t \equiv G = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \phi_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

and $\theta_t = (X_t, \ldots, X_{t-3})'$, $h_t = (1, 0, 0, 0)'$, $w_t = (\varepsilon_t, 0, \ldots, 0)'$, $n_t = 0$. Thus, any AR(p) model has a state-space representation, and so has any ARMA model.

The benefits of state-space models

- Most time-series models can be represented in their state-space form.
- State-space models are more a technique of representing models than a separate class of models.
- The basic form may be convenient, and it allows many generalizations beyond standard time-series models.

A 'structural' model according to Harvey

The simplest of the unobserved-components (UC) models due to Harvey has two state variables μ and β:

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

With white-noise input, X_t is a special I(2) or ARIMA$(p, 2, q)$ process. UC adepts claim that the different parameterization—variances of errors instead of ARMA coefficients—is more 'natural'. UC models sometimes perform surprisingly well in forecasting economic data.