Nonlinear Models

by Kevin Kotzé

In econometric time series analysis we usually try to make use of the information that we have at our disposal to describe a relationship between various variables. Such analyses would usually involved time series that extend over quite a long period of time and as a result of which it may be prudent to incorporate certain changes that have occurred to the parameter estimates of the DGP over the said period. The application of nonlinear time series analysis seeks to incorporate changes to the parameters estimates and the use of regime switching models are particularly relevant to cases where we seek to obtain meaningful parameter estimates from a DGP that is subject to changes that are either observed or unobserved.

These regime switching models seek to incorporate that bit of information that characterizes the dynamic state dependent behaviour of economic or financial variables. For example, the speed of dynamic adjustments of an economic process during a recession may differ from the speed in expansionary periods. If this bit of information is not included in the model and we assume that expansions are more common than recessions then estimated models will treat the observations for recessionary periods as outliers, producing incorrect parameter estimates for the beta parameters and the extremely large distributions for the error terms (see Figure 1).

A further example would occur in cases that consider a model for stock market returns, for which large negative returns are more common than large positive returns (i.e. skewness < 0) and large absolute returns are frequently observed (i.e. kurtosis > 3). If we choose to ignore the apparent nonlinearities in the DGP for this model we would then seek to force an inappropriate joint distribution for the beta parameters (see Figure 1).

These regime switching models could also be used to identify the specific point in time where the DGP moves from one regime to another. For example, we could use a regime switching model to identify changes in the business cycle by specifying the indicator variable as a parameter that is to be estimated. Other applications for these models are numerous and could be structured to describe multiple equilibria or instances where there is no equilibrium at all (i.e. limit cycle).[1]

This lecture will not consider several other nonlinear models such as Generalized Autoregressive models, Bilinear models, Flexible Fourier Form models, Project Pursuit models, along with several other others; as it may be argued that the application of these models is not as widespread in economic and financial fields.[2]

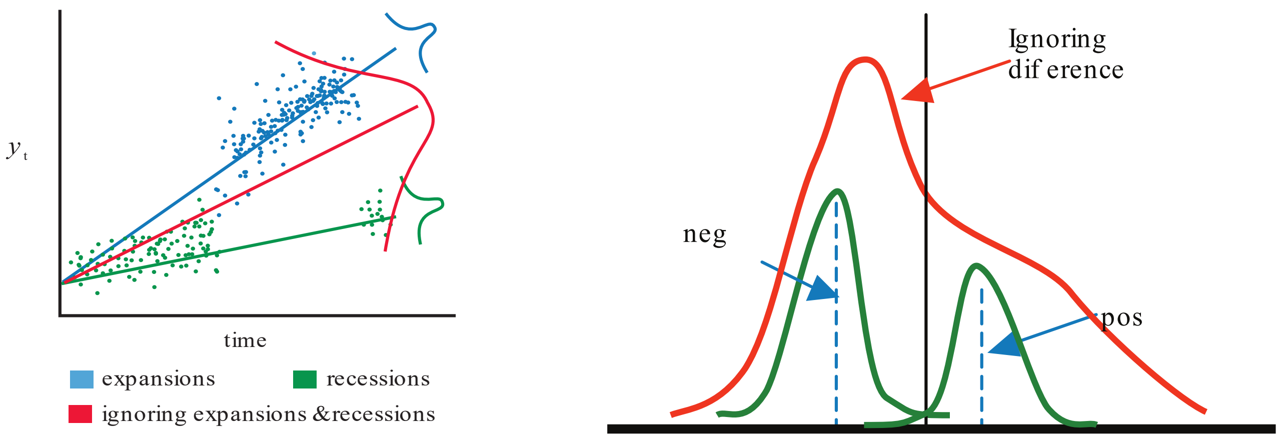


Figure 1: **Distributions of variables with different regimes**

1 The use of dummy variables

For the sake of completeness a brief review of time series models that make use of dummy variables is included below. These models assume that the sample can be easily split into separate groups (regimes) where the parameters are constant within the groups (regimes) but differ between groups. It is important to note that in dummy variable time series models the identification of the groups (regimes) is known with certainty in advance, which would imply that the regime switching process is deterministic.

The use of the promotional dummy variable(s) makes it possible for the researcher to remove that part of the residual that was generated as a result of the promotion and as a result we are often able to estimate more accurate parameters.

In many instances we are not provided with an indicator variable that could be used to make distinctions between the respective regimes and in such circumstances we may seek to identify structural breaks in the time series. To ascertain whether the parameters have changed at these break points one would usually make use of tests such as the Chow break test.[3] This test is performed by splitting the time series sample into groups (e.g. one group of n1 observations and another with n2 observations) before estimating respective regressions for each of the subsets (e.g. $y_1 = \phi_1 x_1 + \varepsilon_1$, and $y_2 = \phi_2 x_2 + \varepsilon_2$). The null hypothesis of constant coefficients is then given by $H_0 : \phi_1 = \phi_2$, which may be tested against the alternative that $\phi_1 \neq \phi_2$ by means of the F-test.

Furthermore one could make use of the CUSUM test to ascertain whether there is possible instability in the parameters within the respective sample.[4] To perform this test one needs to calculate the forecast errors, f_t before deriving the recursive residuals, ω_t , as follows:

$$
\begin{aligned}
f_t &= y_t - \phi_{t-1} x_t \\
\omega_t &= \frac{f_t}{\sqrt{1 + x_t' \left(X_{t-1}' X_{t-1} \right)^{-1} x_t}} \\
\text{where } \ \sigma_{ft}^2 &= \sigma^2 \left[1 + x_t' \left(X_{t-1}' X_{t-1} \right)^{-1} x_t \right]
\end{aligned}
$$

To complete the CUSUM test which is based on the cumulative sum of residuals, find:

$$W_r = \sum_{t=k+1}^{r} \frac{\omega_t}{s} \;\; r = K+1, \ldots, n.$$

Where s^2 is the estimate of σ^2 in the model $y_t = \phi x + \varepsilon$ over the full data sample using all n observations. If the parameters are not susceptible to significant variation then the terms ω_t / s are independent with distribution $\mathcal{N}(0,1)$ so that W_r is approximately distributed as $\mathcal{N}(0, r-k)$.

2 The basic regime switching model

In contrast with the above, the following sections assume that the regime switching process is described by a stochastic process where the future state of the DGP is not known with certainty. This implies that before we are able to complete the estimation of the model we would need to specify the stochastic regime switching process with the aid of certain statistical techniques.

Although there are many ways that one could choose to specify the stochastic regime switching process which may roughly be differentiated as follows;

- where the regime is determined by an observable variable and as such the regimes of the past and the present are known with certainty (i.e. TAR or STAR models), or
- where the future regime is characterized by an unobserved stochastic process and as such the regimes are not known with certainty although probabilities may be assigned to them (i.e. MSW model)

By way of an explanation of the representation of a general regime switching model consider a DGP that is described by two regimes, a constant, ϕ_0, and a set of explanatory variables, x_t . We may then choose to define a linear expression for regime 1 by;

$$y_t = \phi_{0,1} + \phi_{1,1} x_t + \varepsilon_{1t}$$

and the linear expression for regime 2 may be described as;

$$y_t = \phi_{0,2} + \phi_{1,2} x_t + \varepsilon_{2t}$$

Although each of these expressions is linear, the possibility of regime switching means that the model has nonlinear features. This expression may also be written more compactly as;

$$y_t = \begin{cases} \phi_{0,1} + \phi_{1,1} x_t + \varepsilon_t & \text{in regime one} \\ \phi_{0,2} + \phi_{1,2} x_t + \varepsilon_t & \text{in regime two} \end{cases}$$

where $\varepsilon_t \sim (0, \sigma_i^2)$ in regime $i, i = 1, 2$.

It is important to note that in contrast to the models that use dummy variables to account for the unexplained part of the model that is influenced by a structural change, this basic regime switching model attempts to model each of the regimes explicitly. This is of particular relevance when interpreting the model parameters and is also extremely important when estimating forecasting models, as they provide a means of describing future changes in the current regime.

Furthermore, it should also be noted that the variance of the error term in regime one need not equal the variance in regime two, which may be a useful feature when modelling volatility.

3 Deterministic regime switching models

3.1 Threshold Autoregressive (TAR) models

The TAR model assumes that the regime that occurs at time t is characterized by the value of an observable variable, q_t, relative to the value of a threshold, c . Therefore, in a two regime model;

$$y_t = \begin{cases} \phi_{0,1} + \phi_{1,1} x_t + \varepsilon_t & \text{if} \;\; q_t \leq c \\ \phi_{0,2} + \phi_{1,2} x_t + \varepsilon_t & \text{if} \;\; q_t > c \end{cases}$$

where $\varepsilon_t \sim (0, \sigma_i^2)$ in regime $i, i = 1, 2$.

To estimate a model for this system simply separate the data into each regime and use either OLS, a likelihood function or one of the many other methods of parameterisation to find the parameters $\phi_1, \phi_2, \sigma_1^2, \sigma_2^2$.

Example: Threshold Autoregressive model: (Shen and Hakes 1995)

Shen and Hakes use a TAR model to investigate the reaction function of the Taiwanese central bank to determine whether it did in fact pursue its number one priority of price stability, whilst seeking to increase output growth when price stability was not threatened. They used inflation as the threshold variable and the targeted inflation rate as the threshold to consider four regimes during periods of; no inflation, low inflation, moderate inflation, and high inflation.

Before formulating the TAR model they firstly proceeded with an F-test for nonlinearity and rejected the null of linearity at the 5% level. They then proceeded with the identification of the locations for the thresholds and estimate a reaction function (that takes the general form of an open economy Taylor rule).

They find that during periods of no inflation the central bank appears to pursue output growth and low inflation, during periods of low inflation it responds to output (with no response to inflation), and during periods of moderate and high inflation it responds only to inflation and not output growth. Thus it appears as if the central banks activities are consistent with their stated objectives in that it only targets output growth when price stability is not threatened.

3.2 Self Extracting Threshold Autoregressive (SETAR) models

The SETAR model could be regarded as a special case of the TAR model as it assumes that the observable variable, q_t, is a lagged value of the series itself. In this case, where $q_t = y_{t-d}$ for integers $d > 0$, the regime is determined by the lagged value of the variable that is to be explained.

Therefore, where $d = 1$, a two regime SETAR model that may be characterized by the AR(1) process $y_t = \phi_i^\prime y_{t-1} + \varepsilon_t$ during each regime may be given as;

$$y_t = \begin{cases} \phi_{1,1} y_{t-1} + \varepsilon_t & \text{if} \;\; y_{t-1} \leq c \\ \phi_{1,2} y_{t-1} + \varepsilon_t & \text{if} \;\; y_{t-1} > c \end{cases}$$

where $\varepsilon_t \sim (0, \sigma_i^2)$ in regime $i, i = 1, 2$.

In this model the shocks, ε_{1t} and ε_{2t} are responsible for the regime switching. For example, if $y_{t-1} > c$, then the subsequent values of the sequence will tend to decay towards zero at a rate of ϕ_1. However, a negative realization of ε_{1t} can cause y_t to fall by such an extent that it lies below the threshold.

Now if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the model may be written as;

$$y_t = (\phi_{0,1} + \phi_{1,1} y_{t-1})(1 - I\left[y_{t-1} > c\right]) + (\phi_{0,2} + \phi_{1,2} y_{t-1})(I\left[y_{t-1} > c\right]) + \varepsilon_t$$

where $I[\cdot]$ is an indicator function with $I[\cdot] = 1$ if event 1 occurs and $I[\cdot] = 0$ otherwise.

Even though these basic threshold models appear to be relatively simple, they provide a great deal of flexibility. To see how they have been applied to good effect in the literature, consider the article of Shen and Hakes (1995).

3.3 Smooth Transition Autoregressive (STAR) models

The STAR model allows for a more gradual transition between the two regimes. To do so, the discontinuous indicator function, $I[\cdot]$, is replaced by a continuous function $G[q_t; \gamma, c]$ that changes smoothly from 0 to 1 as q_t increases.

Therefore, the resulting STAR model may be expressed as

$$y_t = \phi_1 x_t (1 - G\left[q_t; \gamma, c\right]) + \phi_2 x_t (G\left[q_t; \gamma, c\right]) + \varepsilon_t$$

where $\varepsilon_t \sim (0, \sigma^2)$ and γ is the smoothing parameter in the continuous function.

A popular choice for the transition mechanism in $G\left[q_t; \gamma, c\right]$ is the logistic function, such that;

$$G\left[q_t; \gamma, c\right] = \frac{1}{1 + \exp(-\gamma[q_t - c])}$$

where $\gamma > 0$.

Examples of the above logistic function for various values of the smoothness parameter, γ , which determines the speed of the parameter switches are been given below, where $q_t = y_{t-1}$ and $c = 0$.

In this instance, it should be noted that:

- for very small values of q_t, $G\left[q_t; \gamma, c\right] \approx 0$ and $y_t \approx \phi_1 y_{t-1} + \varepsilon_t$
- for very large values of q_t, $G\left[q_t; \gamma, c\right] \approx 1$ and $y_t \approx \phi_2 y_{t-1} + \varepsilon_t$
- for intermediate values of q_t, the parameters in the linear model linking y_t and x_t are a linear combination of ϕ_1 and ϕ_2, with the weights determined by the value of the transition function

In addition, it should also be noted that the STAR model contains both TAR and linear models as special cases:

- As $y \rightarrow \infty$, the change in the logistic function becomes almost instantaneous such that GI. This would imply that the STAR model represents a TAR model.
- As $y \rightarrow 0$, the logistic function approaches a constant for all values of q_t such that $G\left[q_t; \gamma, c\right] \rightarrow 0.5$. This would imply that the STAR model represents a linear model as there is very little change in the parameters.

Although the logistic function is the most popular choice for the transition mechanism other functions, such as the exponential function, have also been used to good effect in the literature.

3.3.1 Parameter estimation

Although it is possible to estimate the parameters of a STAR model with an OLS procedure it is more common for one to use a nonlinear estimating procedure such as Nonlinear Least Squares (NLS) that also allow for models that are nonlinear in the parameters. This method of parameter estimation is a form of optimization technique, which makes use of certain algorithms that look to minimize an objective function with the aid of an iterative search process that is discussed in Davidson and MacKinnon (2004), Greene (2003), Hayashi (2000) and Judge et al. (1985).

To understand the principles that underpin this procedure consider the general representation of the regression model;

$$y_t = f(x_t, \phi) + \varepsilon_t \qquad t = 1, 2, \ldots, n$$

where f is a general function of the explanatory variables x_t and the parameters ϕ. If the model is linear in parameters the derivatives of f do not depend upon ϕ. This allows us to write the general regression model as;

$$\begin{aligned} y_t &= E(y_t|\mathbf{x}_t) + \varepsilon_t \\ &= \phi_1 x_{t1} + \phi_2 x_{t2} + \ldots + \phi_p x_{tp} + \varepsilon_t \end{aligned}$$

where $t = 1, 2, \ldots, n$ and the corresponding matrix $\mathbf{x}_t$ represents the independent explanatory variables $\mathbf{x}_{ti}$. Following the standard procedures for a linear regression analysis, the estimated parameters and residual sum of squares are calculated as;

$$\begin{aligned} \hat{\phi} &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \\ S(\hat{\phi}) &= \sum_{t=1}^{n} \left[y_t - \hat{\phi}_1 x_{t1} - \hat{\phi}_2 x_{t2} - \ldots - \hat{\phi}_p x_{tp} \right]^2 \end{aligned}$$

Least squares estimation involves the selection of the parameter values that minimize the sum of square residuals. As an alternative, the parameter estimates could also be derived from an iterative search technique that calculates the residual sum of square as;

$$S(\hat{\beta}) = \sum_{t=1}^{n} \left[y_t - \tilde{\phi}_1 x_{t1} - \tilde{\phi}_2 x_{t2} - \ldots - \tilde{\phi}_p x_{tp} - (\hat{\phi}_1 - \tilde{\phi}_1) x_{t1} - (\hat{\phi}_2 - \tilde{\phi}_2) x_{t2} - \ldots - (\hat{\phi}_p - \tilde{\phi}_p) x_{tp} \right]^2$$

where the initial guess for the value of the coefficients in matrix ϕ is expressed as $\tilde{\phi}$. Now consider a general expression that may be used to represent a model that is nonlinear in parameters;

$$y_t = f(\mathbf{x_t}, \psi) + \varepsilon_t \qquad t = 1, 2, \ldots, n$$

where $\psi = (\psi_1, \psi_2, \ldots, \psi_p)$ represents the parameters in the nonlinear regression, such that $\psi = (\phi; G\left[q_t; \gamma, c\right])$. In this instance, we assume that the derivatives of f are functions of ϕ, which suggests that the model has nonlinear parameters. The existence of this relationship implies that it would not be possible to derive a closed form solution for the parameter estimates. This means that the iterative search technique would need to be employed to find the minimum residual sum of squares. Given a vector that contains an initial guess of the ψ coefficients, the residuals may be calculated as $\tilde{\varepsilon} = (\mathbf{y} - \tilde{\mathbf{y}})$. This implies that the residual sum of squares may be determined as;

$$S(\tilde{\psi}) = \tilde{\varepsilon}'\tilde{\varepsilon} = (\mathbf{y} - \tilde{\mathbf{y}})'(\mathbf{y} - \tilde{\mathbf{y}})$$

where $\tilde{\mathbf{y}} = f(\tilde{\psi})$ is a vector of predicted values that is obtained by replacing the unknown parameters by the initial guess values. To approximate the model $f(\mathbf{x}_t, \tilde{\psi})$, the first order Taylor expansion around the initial value of $\tilde{\psi}$ is used to find,

$$f(\tilde{\psi} = f(\tilde{\psi})\mathbf{x}_{\tilde{\psi}}\delta$$

where $\delta = (\psi - \tilde{\psi})$ and $\mathbf{x}_{\tilde{\psi}} = x_{tj}$ is a $n \times p$ matrix of the partial derivatives at $\tilde{\psi}$ in the above linear approximation. To determine the value of γ it is then possible to calculate;

$$\delta = \left(\mathbf{x}'_{\tilde{\psi}}\mathbf{x}_{\tilde{\psi}}\right)^{-1}\mathbf{x}_{\tilde{\psi}}\tilde{\varepsilon} = (\delta_1, \delta_2, \ldots, \delta_p)$$

In a nonlinear model, the values in the matrix $\mathbf{x}_{\tilde{\psi}}$ will change from one iteration to another. Therefore, the updated least squares estimates $(\hat{\psi} - \tilde{\psi} - \delta)$ would need to be calculated after each iteration. These values may then be used to replace the values that were used as the initial guess, and further iterations will continue until the difference between successive parameter vectors, δ, is small enough to assume that the a priori convergence criteria has been satisfied.

Hence, this procedure determines the least square estimates that minimize the residual sum of squares, where;

$$\tilde{\psi} = \min \sum_{t=1}^{n} (y_t - f(\mathbf{x}_t, \psi))^2$$

To reduce the time taken to achieve convergence, procedures such as the Gauss-Newton and Levenburg-Marquardt algorithms are frequently used to derive impressive results.

3.3.2 Testing: regime-switching nonlinearity

Perhaps one of the most important questions that we need to answer before estimating a regime switching model is whether the additional regime adds significantly to the explanation of the dynamic behaviour of the time series.

To test for the parameter variation as suggested by a inclusion of multiple regimes one may wish to conduct a Likelihood Ratio (LR) test which is based on the loss of log-likelihood following the imposition of certain restrictions that are placed on the model (i.e. in this case the restriction is linearity). This test is unfortunately rather cumbersome as it requires the estimation of both models to obtain the parameters which are used to calculate the log-likelihood’s of the respective nonlinear and linear models. If the test statistic is sufficiently large, relative to the p-value from the χ^2 distribution then the null hypothesis is rejected.

As an alternative one may wish to make use of a Lagrange Multiplier (LM) test that is used to test for a specific type of nonlinearity. If we assume that the variance of the error term σ^2 is constant and $f(\cdot)$ is the functional form of the model that has parameters ϕ then the test is conducted as follows:

- estimate the linear portion of the model to get the residuals, ε_t
- obtain the partial derivatives $\partial f(\cdot)/\partial\phi$ evaluated under the null of linearity and estimate an auxiliary regression by regressing ε_t on the partial derivatives
- if the calculated (uncentered) R^2 exceeds the critical value of from a χ^2 squared table, reject the null of linearity and accept the alternative

Although this test is relatively simple to perform there is a problem in applying it to certain nonlinear models since the parameters that define the functional form of the nonlinear model must be identified under the null of linearity. This however is not necessarily a problem for the STAR model as the linear specification of the model is nested within its functional form (i.e. when $\gamma = 0$ we know that $G\left[q_t; \gamma, c\right] \rightarrow 0.5$).

To perform this test on a STAR model we replace the transition function $G\left[q_t; \gamma, c\right]$ with a third order Taylor expansion around $\gamma = 0$. This process yields the auxiliary model:

$$y_t = \beta_{0,0} + \beta_0' \tilde{x}_t + \beta_1' \tilde{y}_{t-1} + \beta_2' \tilde{y}_{t-2}^2 + \beta_3' \tilde{y}_{t-3}^3 + \varepsilon_t$$

where $\beta_{0,0}$ and the β_j , $j = 1, 2, 3$ are functions of the parameters ϕ_1, ϕ_2, γ and c. Inspection of the exact relationships demonstrates that the null $H_0 : \gamma = 0$ now corresponds to $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, which can be tested with a standard LM test using a χ^2 distribution with $3p$ degrees of freedom.

In small samples we would usually make use of an F-test statistic as it has better size and power properties. It would be computed as:

- estimate the model under the null hypothesis of linearity by regressing y_t on x_t. Compute the residuals $\tilde{\varepsilon}_t$ and the sum of squared residuals $SSR_0 = \sum_{t=1}^{n} \tilde{\varepsilon}_t$
- estimate the auxiliary regression of $\tilde{\varepsilon_t}$ on x_t and $x_t y_t$ and compute the sum of squared residuals from this regression, SSR_1
- the LM test statistic can then calculated as:

$$LM = \frac{(SSR_0 - SSR_1)/3p}{SSR_1/(n - 4p - 1)}$$

which is approximately F-distributed with $3p$ and $n - 4p - 1$ degrees of freedom under the null hypothesis.

3.3.3 Testing: diagnostics

Unfortunately not all of the test statistics that are used in the study of linear models are applicable to the residuals of nonlinear models. We can however still make use of the LM approach to testing diagnostics and some of the test that make use of this method are discussed below:

- Testing for serial correlation:

In the general form of the nonlinear model,

$$y_t = f(\mathbf{x_t}, \psi) + \varepsilon_t \qquad t = 1, 2, \ldots, n$$

where $\psi = (\psi_1, \psi_2, \ldots, \psi_p)$ represents the parameters in the nonlinear regression. An LM test for the qth order serial correlation in ε_t can be obtained as nR^2, where R^2 is the coefficient of determination from the regression of the estimated ε_t on z_t, which represents the expression $\partial f(\mathbf{x_t}, \hat{\psi}_\mathbf{t} / \partial \psi_\mathbf{t}$ and q lagged residuals $\varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$. The resulting statistic is χ^2 distributed with q degrees of freedom

- Testing for remaining nonlinearity:

Various tests have been developed for remaining nonlinearity. These tests also take the form of LM statistics where we may test the null of whether a 2 regime STAR model adequately capture the nonlinearity (i.e. against the alternative of whether a 3 regime STAR model should be used).[5]

- Testing parameter constancy

Similar tests have also been developed for parameter constancy, which seek to investigate whether we would need to include time varying parameters. The process basically involves testing the hypothesis $\gamma_2 = 0$ against the alternative of smoothly changing parameters.

Example: Smooth Transition Autoregressive model: Franses and Dijk (2000)

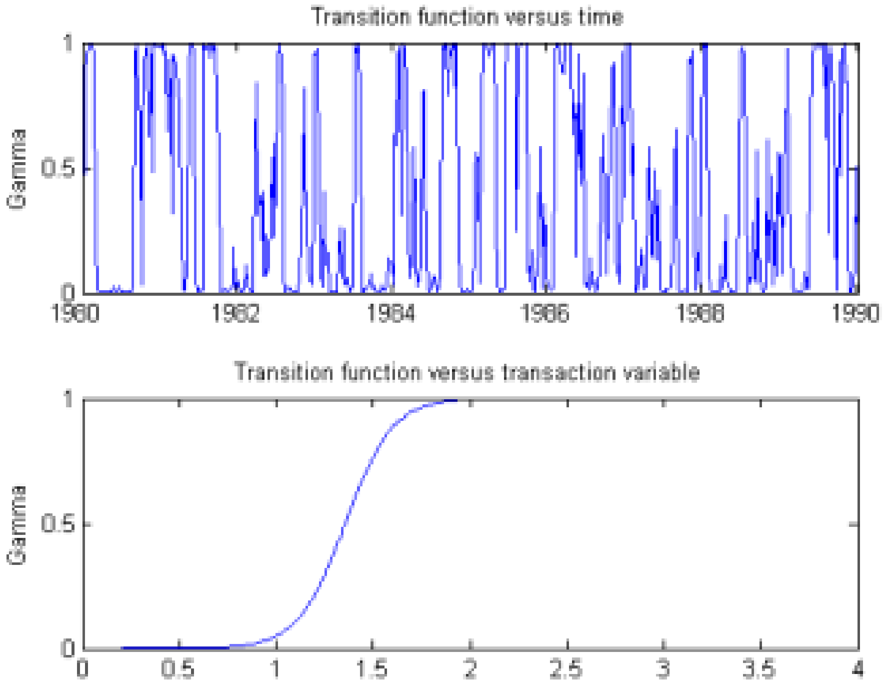
Franses and Dijk (2000) consider a STAR model that sought to describe the behaviour of an exchange rate (Dutch guilder). After rejecting the null of linearity they estimated a two regime STAR model for an AR(2) process where the threshold variable is the average of exchange rate for the last 4 weeks.

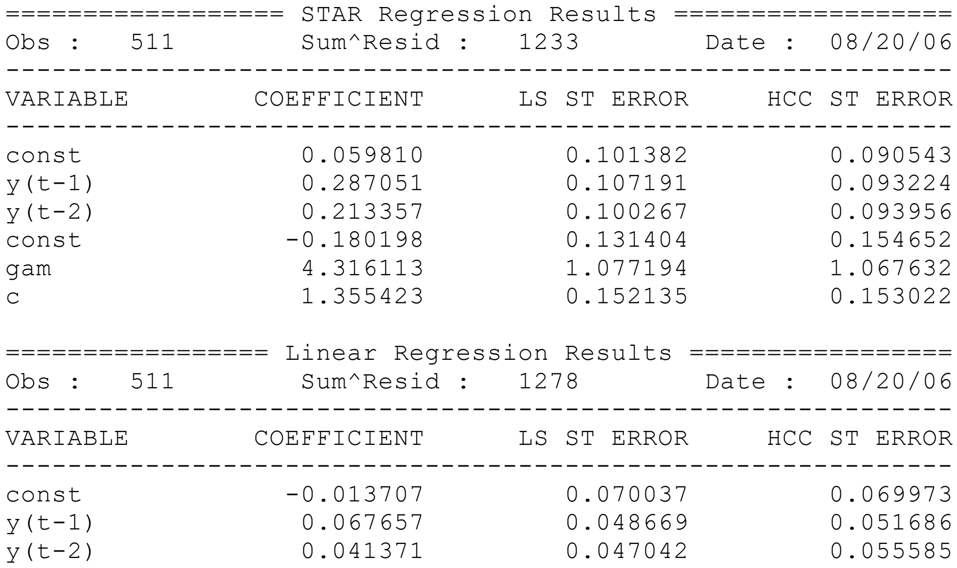
In this example, uncertainty surrounded the calibration of the threshold and the value of the smoothing parameter, gamma. Hence, an optimization procedure was invoked to test various values for these parameters to find those values that would ultimately minimize the sum of square errors in the final model.

To derive starting values for the parameters in the optimization process a simplified grid search technique was used where the parameter space is split into incremental sections over a specified range (as per, Hamilton (1994). The parameter values that generated the smallest residual are then used as starting values in the optimization procedure.

Once suitable values for the gamma and threshold parameters were obtained the remaining parameters in each regime were estimated with OLS to produce the results that have been reproduced with the aid of my program `lstar_est.m`.

As seen by the value of the gamma parameter the graph for the logistic function is relatively smooth, as the system moves from one regime to the other. Unfortunately, the autoregressive parameters in the regime corresponding to $G\left[q_t; \gamma, c\right] = 1$ were not significant so the model had to be re-estimated after deleting these parameters and that is why the values for y_{t-1} and y_{t-2} were not reported in the alternate regime.





```
================== STAR Regression Results ==================
Obs :   511        Sum^Resid :   1233        Date :  08/20/06
-------------------------------------------------------------
VARIABLE        COEFFICIENT      LS ST ERROR     HCC ST ERROR
-------------------------------------------------------------
const              0.059810         0.101382         0.090543
y(t-1)             0.287051         0.107191         0.093224
y(t-2)             0.213357         0.100267         0.093956
const             -0.180198         0.131404         0.154652
gam                4.316113         1.077194         1.067632
c                  1.355423         0.152135         0.153022

================ Linear Regression Results =================
Obs :   511        Sum^Resid :   1278        Date :  08/20/06
-------------------------------------------------------------
VARIABLE        COEFFICIENT      LS ST ERROR     HCC ST ERROR
-------------------------------------------------------------
const             -0.013707         0.070037         0.069973
y(t-1)             0.067657         0.048669         0.051686
y(t-2)             0.041371         0.047042         0.055585
```

Figure 2: **Smooth Transition Autoregressive model**

4 Specification procedures for regime switching models

Granger and Teräsvirta (1993) advocate:

"Building models of nonlinear relationships are inherently more difficult than linear ones. There are more possibilities, many more parameters and thus more mistakes can be made. It is suggested that a strategy be applied when attempting such modelling involving testing for linearity, considering just a few model types of parsimonious form and performing post-sample evaluation of the models compared to a linear one. The strategy proposed is a 'simple-to-general' one and the application of a heteroscedasticity correction is not recommended."

With this in mind, it is suggested that the specification procedure for nonlinear models should be as follows;

- Specify a linear model to describe y_t in terms of x_t.

After specifying a linear model it would be a good idea to test for the presence of nonlinearity. Various portmanteau, RESET and Macleod-Li tests may be used identify nonlinearity, however they are not able to provide any guidance as to the nature of the nonlinearity.

- Test the null hypothesis of linearity against the alternative of TAR, STAR, MSW, or ANN nonlinearity.

If it is assumed that the relationship has nonlinear properties we would need to estimate a nonlinear model to obtain values for the parameters which could then be used to construct a Lagrange Multiplier test to ascertain whether the model adequately captures the nonlinear relationship between the variables.

To conduct this test we often;

- estimate the linear portion of the model to get the residuals ε_t
- obtain the partial derivatives $\partial f / \partial \alpha$ evaluated under the null of linearity and estimate the auxiliary regression by regressing et on these partial derivatives
- if the calculated value of the TR^2 exceeds the critical value from the χ^2 table (with d.o.f. equal to the number of regressors in b) then reject the null of linearity and accept the alternative

It is also worth noting that when testing the null of linearity against the alternative of a particular nonlinearity, the existence of certain unidentified nuisance parameters (c and γ) would suggest that the statistics are non-standard. This would imply that critical values have to be determined by means of simulations methods.

- Evaluate the model using diagnostic tests.

Standard diagnostic tests would include various LM type tests for serial correlation, parameter consistency, heteroscedasticity, and omitted variables.

- Modify the model if necessary.
- Use the model for descriptive or forecasting purposes.

5 Markov Switching Models

5.1 Introduction

In certain instances, we are not always able to observe a reliable variable that we could use as the regime indicator, q_t . In these instances, the regime that occurs at time t is not deterministic and as a result we would then need to describe the unobserved process, S_t , which causes the system to change from one regime to another by a probabilistic model.

If we were only to have two regimes then we may choose to express the model as;

$$y_t = \begin{cases} \phi_{0,1} + \phi_{1,1} x_t + \varepsilon_t & \text{if } \ S_t = 0 \\ \phi_{0,2} + \phi_{1,2} x_t + \varepsilon_t & \text{if } \ S_t = 1 \end{cases}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_{S_t})$ in regime $i, i = 1, 2$.

5.2 Transition probabilities and the Markov chain

In the Markov Switching model, the unobserved process is characterised by a first order Markov process, where the determination of the current regime depends upon the regime from one period ago. This would imply that the probability of a change in the current regime depends on past regimes, but only in so much as the past regimes are reflected in the most recent regime and the overall probability that the regime will change from one to another.

Hence, it is assumed that the conditional distribution of $x_{t,2}$ at $t, 2$ will only depend on $x_{t,1}$ and the respective probability p_{ij}, and not on the value of x at an earlier point in time.

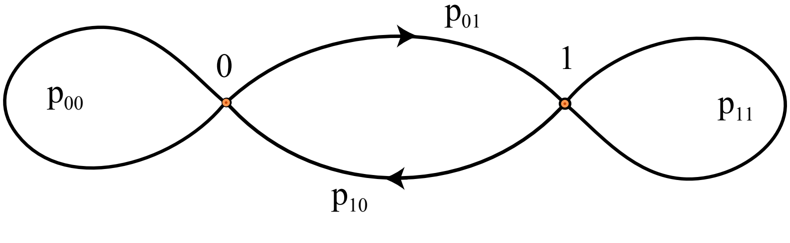


Figure 3: **Markov Switching Transition Probabilities**

The above diagram represents a two regime Markov chain, where it is assumed that the system will either be in regime 0 or regime 1. At any transition time t there is a probability p_{ij} that the system, if in regime i will change to regime j (where $i, j = 1, 0$).

Thus p_{01} is the probability of a change in the system from regime 0 to regime 1, and p_{11} is the probability of the system remaining in regime 1 if it is already there. Hence,

$$\begin{aligned}
&P(s_t = 0|s_{t-1} = 0) = p_{00} \\
&P(s_t = 1|s_{t-1} = 0) = p_{01} = 1 - p_{00} \\
&P(s_t = 0|s_{t-1} = 1) = p_{10} = 1 - p_{11} \\
&P(s_t = 1|s_{t-1} = 1) = p_{11}
\end{aligned}$$

{where $p_{10} + p_{11} = 1$ and $p_{00} + p_{01} = 1$

You will note from the above expressions that the transitional probabilities are completely defined by p_{00} and p_{11}, since $p_{01} = 1 - p_{00}$ and $p_{10} = 1 - p_{11}$. Therefore, if $p_{10} = p_{01} = 1$ and $p_{00} = p_{11} = 0$, then the system will be in continuously change. Similarly, if $p_{10} = p_{01} = 0$ and $p_{00} = p_{11} = 1$, then the system will never change out of one regime. These parameters are all conditional probabilities as the state of the future regime is dependent upon the previous state.

Although the above cases may be of interest from an explanatory perspective it is important to note that in the majority of instances the value of p_{ij} would be between 0 and 1, thus allowing for varying degrees of regime persistence (as apposed to regime permanence or impermanence). This persistence is facilitated by another special property of Markov chains, whereby it can be shown that the system will approach a stable point (if all the probabilities are nonzero) that may be described by the following unconditional (steady-state) probabilities of being in each of the respective regimes at any point in time;[6]

$$\begin{aligned}
P[S_0] &= \frac{1 - p_{11}}{2 - p_{00} - p_{11}} \quad \text{or} \quad \frac{p_{10}}{p_{01} + p_{10}} \\
P[S_1] &= \frac{1 - p_{00}}{2 - p_{00} - p_{11}} \quad \text{or} \quad \frac{p_{01}}{p_{01} + p_{10}}
\end{aligned}$$

In the exposition of this model, the exogenous transitional probabilities are time invariant, which implies that although the expected duration of expansions and recessions can differ (i.e. an economic decline may last for 2,6,12, etc. quarters) they are forced to be constant over time (i.e. the probability of a change does not change over time). These models are therefore often referred to as “fixed transition probability Markov Switching models”.

5.3 Time varying transition probabilities

In certain instances we may want to allow for the transitional probabilities to vary over time. For example, when modelling business cycles or exchange rate we may want to assume that as the economy exits a relatively deep recession and enters a relatively robust recovery period, it is less likely to fall back into a recession. To facilitate the incorporation of these time varying transition probabilities (TVTP) properties we would either need to make p_{ij} a function of duration, or as an alternative we could make p_{ij} a function of another variable.

Essentially, the specification of the transitional probabilities would then need to allow for the incorporation of the additional information that is provided by the expected duration / variable. As such, they would be specified as;

$$
\begin{aligned}
P(s_t = 0|s_{t-1} = 0, \psi_t) = p_{00}(\psi_t) \\
P(s_t = 1|s_{t-1} = 0, \psi_t) = p_{01}(\psi_t) \\
P(s_t = 0|s_{t-1} = 1, \psi_t) = p_{10}(\psi_t) \\
P(s_t = 1|s_{t-1} = 1, \psi_t) = p_{11}(\psi_t)
\end{aligned}
$$

where ψ_t represents the information set upon which the evolution of the unobserved regime will depend, and $1 > p_{ij} \geq 0, p_{i0}(\psi_t) + p_{i1}(\psi_t) = 1$ and $p_{0j}(\psi_t) + p_{1j}(\psi_t) = 1$.

To incorporate these TVTP into the model we would normally use a logistic function to allow for the transition probabilities to vary over time. In the case where the transitional probabilities are dependant on the value of an exogenous variable, z_t, the transitional probabilities may then be expressed as;

$$
\begin{aligned}
p_{00} = P(S - t = 0|S_{t-1} = 0) = \frac{\exp(\alpha_0 + \beta_0 z_t)}{(1 + \exp(\alpha_0 + \beta_0 z_t)} \\
p_{11} = P(S - t = 1|S_{t-1} = 1) = \frac{\exp(\alpha_1 + \beta_1 z_t)}{(1 + \exp(\alpha_1 + \beta_1 z_t)}
\end{aligned}
$$

For more information about TVTP, see Filardo (1994).

5.4 Estimation of the model parameters

The estimation procedure for the Markov Switching model is non-standard since it not only seeks to obtain the estimates of the parameters in the different regimes along with the probabilities of transition from one regime to another, but it also seeks to estimate the probability with which each state occurs at each point in time.

To do so it makes use of an iterative algorithm to calculate all of the objects of interest. This procedure is similar in nature to the application of the Kalman filter and was established in Hamilton (1989) which makes use of a filtering algorithm that quantifies the conditional probability of being in each regime for each and every point in time.[7] In essence, the algorithm is responsible for calculating the probability that the process is in regime j at time t given:

- all observations up to time $t - 1$ (i.e. the forecast)
- all observations up to time t (i.e. the inference)
- all observations in the entire sample (i.e. the smoothed inference)

In many regards this algorithm is rather unique in that not only is it responsible for the calculation of the conditional probabilities, but as a by-product, it also estimates the other parameters as well.

The basic setup makes use of an iterative procedure which includes:

- obtain starting values for the model parameters
- compute the smoothed regime probabilities with the aid of the procedures specified in the forecast & inference sections
- combine these estimates with the initial estimates of the transition probabilities to obtain new estimates for the transition probabilities working backwards from n to 1
- calculate values for the remaining ϕ parameters
- iterating this procedure renders a new set of estimates until convergence occurs

5.5 Applications of Markov Switching models

Although Markov Switching models have been used in a variety of different circumstances their macroeconomic applications include,

- Business cycle analysis
    - determination of turning points
    - determination of length of business cycle
    - forecasting turning points
- Appreciation and depreciation regimes in exchange rates
- Different regimes in volatility
- Different regimes in interest rates
- Different political regimes and other institutional characteristics

Example: Hamilton's Markov Switching model for business cycles in the US

A classic example of the Markov Switching model is included in the seminal work of Hamilton (1989). However, as this Econometrica article is a bit of a tough read, it is suggested that you might rather prefer to read the short extract from this paper, ``Illustration - The behaviour of U.S. Real GNP", which is contained from Hamilton (1994).

This model seeks to describe the asymmetry in business fluctuations, where the turning point is treated as a structural event that is inherent to the data generating process. In the construction of the model it is assumed that growth in U.S. Real GNP follows an AR(4) process, which evolves according to a two state Markov switching process for economic expansions / recessions.

Hence;

$$\begin{aligned}(y_t - \mu_{S_t}) = \phi_1(y_{t-1} - \mu_{S_{t-1}}) + \phi_2(y_{t-2} - \mu_{S_{t-2}}) \ldots \\ \ldots + \phi_3(y_{t-3} - \mu_{S_{t-3}}) + \phi_4(y_{t-4} - \mu_{S_{t-4}}) + \varepsilon_t\end{aligned}$$

where $\varepsilon_t = i.i.d. \mathcal{N}(0, \sigma^2_{S_t})$.

In an attempt to replicate the study, I have written the Matlab program `msw_ham.m`, which is contained in the directory `markov1.zip`. This model is responsible for generating the following graph, where the shaded areas represent the NBER dating of recessions.

The above results also suggests that the probability that an expansion will be followed by another quarter of expansion is $p_{00} = 0.9$, so that this regime will persist on average for $1/(1 - p_{00}) = 10$ quarters. Furthermore, it is also worth noting that the probability that a contraction will be followed by another contraction is $p_{11} = 0.76$, such that these episodes are expected to persist for $1/(1 - p_{11}) = 4$ quarters.

In this model, we also note that the average growth rate during an expansion is $\mu_1 = 1.2\%$ per quarter, whilst the average growth rate is $\mu_2 = -0.4\%$ per quarter during a recession.

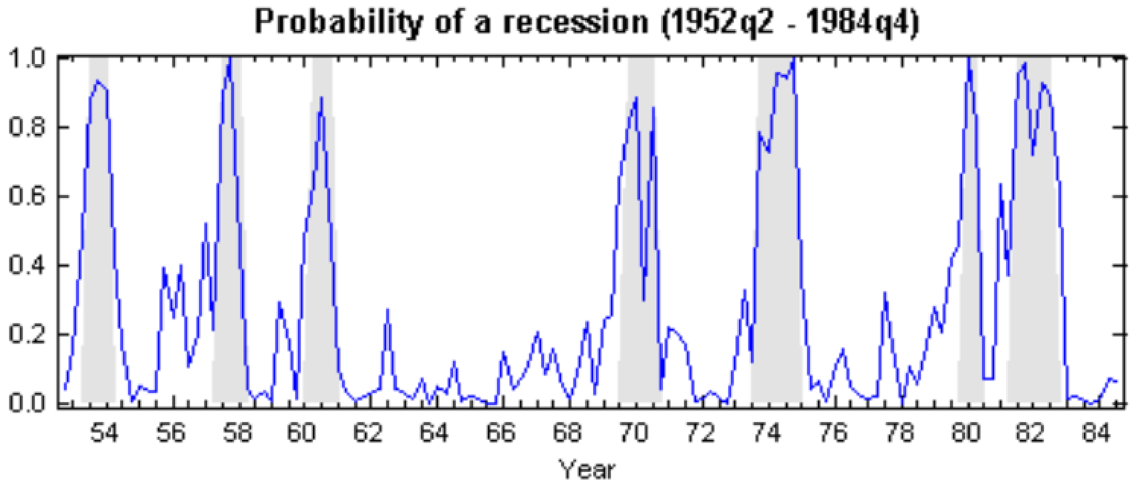


Figure 4: **Markov Switching Application**

Example: Moolman's Markov Switching model for SA

The paper by Moolman (2004) provides an example of the application of Markov Switching models with TVTP, which are influenced by the yield spread (i.e. the difference between similar long-term and short-term instruments).

The rationale for using this financial instrument is that it is said to be related to the business cycle in that if investors perceive that the economy will slow down, they are likely to purchase long-term bonds and sell short-term bonds to obtain consistent coupon payments during the slow down. This would drive up the relative price of long-term bonds, which would cause the yield for these instruments to decline and as a result, the yield curve would flatten out.

Similarly, the expectation hypothesis of the term structure of interest rates suggests that if a central bank were to raise short-term interest rates, it would be expected that economic activity would decline in the near future. Since economic agents would expect that future short-term rates would increase by less than the current rates then the long-term interest rate would increase by the same proportion (if it is assumed that long-term rates are the sum of future short-term rates). Hence the yield curve would flatten out in anticipation of an economic slow down.

The incorporation of the information about where the economy is heading influences the duration of a cycle, such that as the economy exits a relatively deep recession and enters a relatively robust recovery period, it is less likely to fall back into a recession. Therefore, the model which was specified as a two regime first order MSW model that followed as AR(4) process was derived as;

$$\begin{aligned} y_t = \mu_2(1-S_t) + \mu_1 S_t + \phi_1(y_{t-1} - (\mu_2(1-S_{t-1})) + \mu_{S_{t-1}}) + \ldots \\ \ldots + \phi_4(y_{t-4} - (\mu_2(1-S_{t-1})) + \mu_{S_{t-4}}) + \varepsilon_t \end{aligned}$$

The results of this paper may be included in the above graph, where it should be noted that the `false' signals of 1985 and 1994 would not be strictly classified as a change in the cycle since the SARB dating rule for the classification of a change in cycle is that it should persist for at least two quarters.

The above coefficients would suggest that during an economic downturn, the economy would decline by $\mu_1 = -1.06\%$, whilst the economy would be expected to experience an increase of $\mu_2 = 3.74\%$ during an upswing.

Moolman (2004) further showed that the results of this model closely matched those of a logit model which used the official turning points in its estimation and suggested that interest rates should be used as a leading indicator for predicting the business cycle for two quarters into the future. This statement however should have been qualified by stating that it could only be used to a forecast a regime that was expected to persist for two quarters a priori.

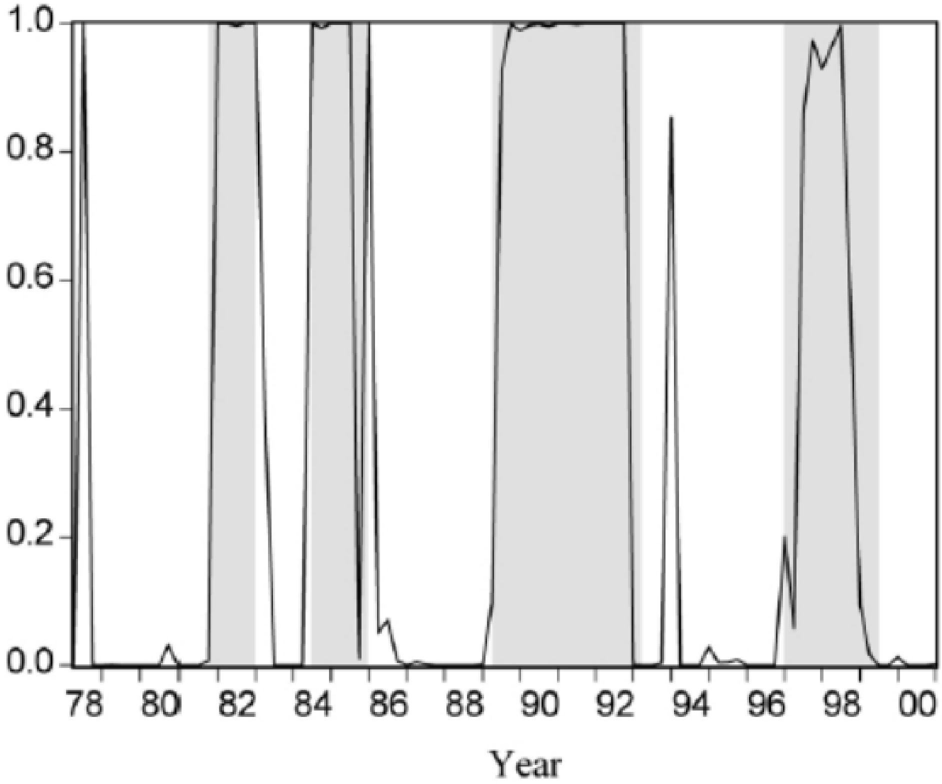


Figure 5: **Markov Switching for South African Business Cycle**

Example: Engle's Markov switching model for forecasting exchange rates

It was argued in Engel and Hamilton (1990) that the US Dollar exchange rate appears to follow long swings as it drifts upwards for a considerable period of time, and then switches to a long period that has a downward drift.

However, although Engel (1994) showed that the Markov Switching model outperforms similar random walk models on in-sample testing it does not necessarily clearly outperform the random walk model or the forward exchange rate in its out-of-sample forecasting ability. This study was relatively comprehensive in that it considered the forecasting of exchange rates in 18 different countries.

Engel (1994) suggested that the model does nevertheless seem to be able to pick up a change in regime fairly early on, however, this result seems to be dependent upon the persistence of a regime for a relatively long period of time.

In addition, these models have also been used to investigate the impact of volatility on a system with the aid of volatility switching autoregressive conditional heteroscedasticity (SWARCH) models.

6 Artificial Neural Networks

ANN models are often regarded as flexible non-parametric models that are able to approximate almost any nonlinear function arbitrarily closely to capture the nonlinear dynamic relationships in the data. However, these models could also be specified as flexible regime switching models and could be interpreted as such.

Before proceeding it is important to note that one of the significant drawbacks of ANN models is that the parameters are difficult (if not impossible) to interpret. This implies that they may only be used for pattern recognition or forecasting purposes. In addition, the superior in sample fit that may be achieved is no guarantee that an ANN model will perform well in out-of-sample forecasting, since it is also able to fit irregular (and unpredictable) noise.

To approximate these models consider what is referred to as a ``single hidden layer feedforward ANN", which takes the form;

$$y_t = \phi_0 + \sum_{j=1}^{1} \beta_j G\left(\gamma_j\left(x_t - c_j\right)\right) + \varepsilon_t$$

In these models we could once again use the logistic function to represent the continuous function $G(\cdot)$;

$$G(z) = \frac{1}{1+\exp(-z)}$$

Now suppose that the parameters $(c_j, j = 1, \ldots, q)$ are such that, $c_1 \leq c_2 \leq \ldots \leq c_q$, and $\gamma_j = +\infty$. This would imply that the logistic function would become a stepwise function, where;

$$G(\gamma_j(x_t - c_j)) = I[x_t > c_j] = \begin{cases} 1 & \text{if } x_t > c_j \\ 0 & \text{if } x_t \leq c_j \end{cases}$$

Then;

$$\hat{g}(x_t) = \begin{cases} \phi_0 & \text{if } x_t \leq c_1 \\ \phi_0 + \beta_1 & \text{if } c_1 \leq x_t \leq c_2 \\ \phi_0 + \beta_1 + \beta_2 & \text{if } c_2 \leq x_t \leq c_3 \\ \vdots & \\ \phi_0 + \beta_1 + \beta_2 + \ldots + \beta_q & \text{if } c_q < x_t \end{cases}$$

If the γ_j , are finite then the logistic functions would change from 0 to 1 only gradually, such that the changes from one level to the next become smooth, which increases the flexibility of the ANN. Although this characteristic would suggest that the model is similar to a STAR model it is important to note that they differ in the following manner;

- Although looks similar to a STAR model it is different;
- STAR: regime is usually determined by 1 lagged value of y_t,
- ANN: regime normally considers p lagged values of y_t.
- STAR: each regime has its own intercept
- ANN: only one intercept is used

- An important difference is ANN models normally use more than one logistic function, which gives it the ability to approximate any nonlinear model arbitrarily closely.

As an alternative, this model could be expressed using the traditional nomenclature;

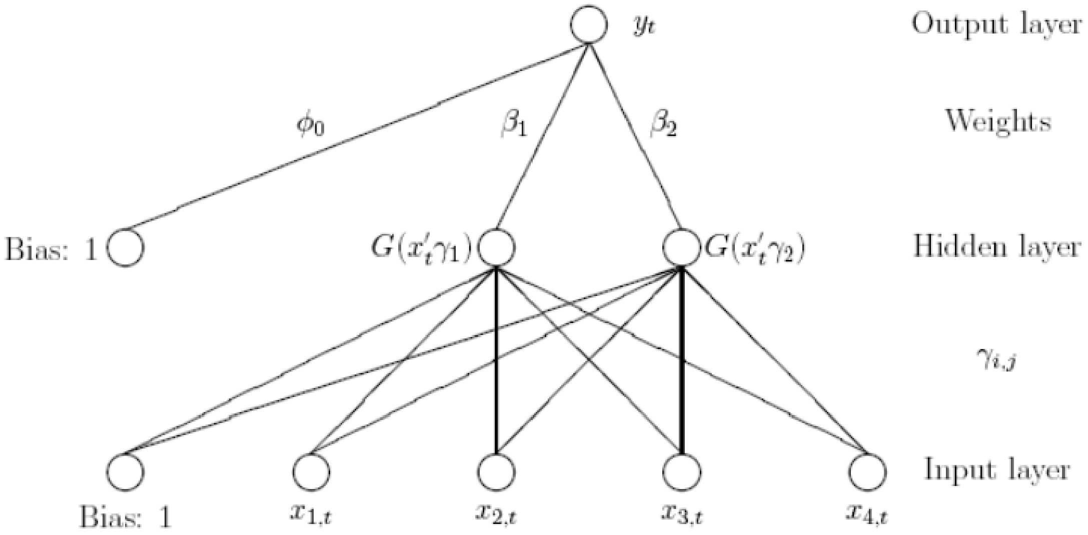
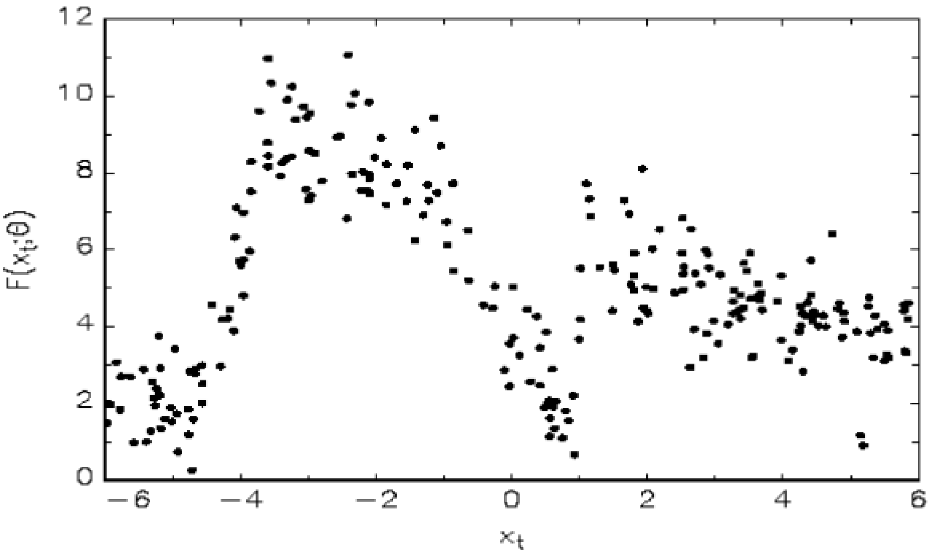


Figure 6: **Traditional nomenclature**

Note the importance of determining the number of q, since models that have a relatively high value for q exhibit a significant amount of flexibility and could be responsible for overfitting.

Example: Simulated ANN model

To show how the models extreme nonlinear features allow it to estimate almost any relationship consider the following scatter plot for a hypothetical relationship between two variables, x_t and y_t .



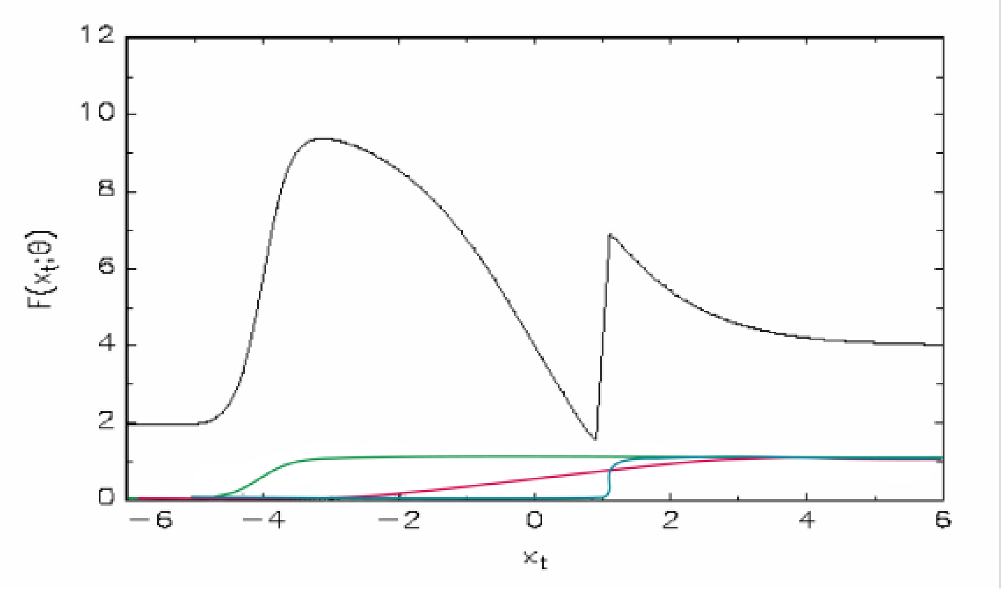


Figure 7: **Observed time series**

Once a suitable ANN model has been estimated one is able to evaluate the model by testing its in-sample fit using traditional methods. In addition, it is also important to conduct various misspecification tests for remaining nonlinearity and parameter constancy. Residual diagnostic tests for serial correlation, heteroscedasticity and normality can also be performed. In addition, it is of extreme importance that this model is subjected to stringent out-of-sample testing, and in this regard the suggestion in Granger and Teräsvirta (1993) that at least 20% of the sample size should be reserved for such testing seems very wise.

7 References

NEURAL NETWORKS, DEEP LEARNING, AND TREE-BASED

In this chapter we introduce some recent developments in high-dimensional statistical analysis that are useful in time series analysis. The methods discussed include neural networks, deep learning, and regression trees. Our discussions focus on analysis of dynamically dependent data and their applications. Deep learning and neural networks have a long history with many successful applications and continue to attract lots of research interest. Readers are referred to Schmidhuber (2015) for an overview and to the online free book by M. Nielsen (2017) for detailed descriptions and computing codes.

8 NEURAL NETWORKS

Neural networks are semi-parametric statistical procedures that have evolved over time with the advancement in computational power and algorithms. They can be used in prediction or in classification.

Figure 4.1. Schematic of a 3-4-2 feedforward neural network. The x nodes are input, the h nodes represent the hidden layer, and the y nodes are output.

In this section, we shall only describe the widely used vanilla feedforward networks. For general networks, readers are referred to textbooks on the topic. See, for instance, Hassoun (2003).

The concept of neural networks originates from models for the human brain, but it has evolved into a powerful statistical method for information processing. Some terminologies of the brain remain in use. Figure 4.1 shows a network diagram for a feedforward 3-4-2 network. The circles in the diagram are referred to as nodes or neurons, and the arrows signify the direction of information flow and the connection between nodes. In this particular network, there are three input variables, denoted by Xs, and two output variables, denoted by Ys. The network has a single hidden layer that has four nodes, denoted by hs. Each input variable is connected to every hidden node, which in turn is connected to every output variable. The name 3-4-2 is used to represent the structure of the network. The output variables Ys can be continuous or discrete. In classification applications, Y is discrete with y, being the probability that the output belongs to the ith class.

In real applications, multiple hidden layers may exist in a network. For simplicity, we only describe the case of a single hidden layer in our discussion. The same idea, however, applies to the general networks.

Suppose that $X = (X_1, \ldots, X_p)'$ is the vector of p input variables, $H = (H_1, \ldots, H_m)'$ is the vector of m hidden nodes, and $Y = (Y_1, \ldots, Y_k)'$ is the vector of k output variables. Thus, we consider a $p - m - k$ feedforward network. An important feature of a given neural network is the way in which the information is processed from input nodes to the hidden nodes, then to the output nodes. Under the statistical framework used, we can write the network as

$$\begin{aligned} H_i &= h(\alpha_{0i} + \alpha_i' X), \quad i = 1, \ldots, m, && (1) \\ L_j &= \beta_{0j} + \beta_j' H, \quad j = 1, \ldots, k, && (2) \\ Y_j(X) &= g_j(L), \quad j = l, \ldots, k, && (3) \end{aligned}$$

where $h(\cdot)$ and $g_j(\cdot)$ are activation functions, α_{0i}, and β_{0j} are real parameters, *i* = ({1i}, ...,_{pi})^$ is a p-dimensional real vector, $\beta_j = (\beta_{1j}, \ldots, \beta_{mj})'$ is an m-dimensional real vector, and $L = (L_1, \ldots, L_k)'$. The linear functions L_j are usedas an intermediate step. The activation function $h(\cdot)$ is often chosen to be the sigmoid function (inverse of the logit function)

$$h(v) = \frac{1}{1 + e^{-v}} \tag{4}$$

The choice of activation function $g_j(\cdot)$ depends on the output variables. If the output variables Y are continuous, then $g_j(L) = L_j$, which is simply the identity function of the jth element of Z. On the other hand, if Y represents classification, then

$$g_j(L) = \frac{e^{L_j}}{\sum_{v=1}^{k} e^{L_v}} \tag{5}$$

which is the *softmax* function satisfying $0 \leq g_j(\cdot) \leq 1$ and $\sum_{j=1}^{k} g_j(L) = 1$, that is, it is a multinomial model. In statistics, α_{ij} and $_{ij} of Equations (4.1)-(4.3) are unknown parameters. In neural networks, they are referred to as *weights*. For a given network, neural network modeling reduces to estimation of the unknown parameters provided that the activation functions are chosen. In this sense, neural networks belong to semi-parametric statistical models. In the special case that $k = 1$ and Y is binary, the activation function in (4.5) may reduce the heaviside function,

$$g_(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases} \tag{6}$$

The resulting simple feedforward neural network is referred to as a perceptron in the literature.

Figure 4.2 shows the sigmoid function $h(\cdot)$ and two of its scaled versions $h(sv)$ with $s = 0.5$ and $s = 5$, respectively. From the plots, it is seen that for small v, $h(v)$ iS approximately linear and $h(sv)$ approaches the $0 - 1$ step function as $s \rightarrow \infty$. In applications, one can use $h(s(v - v_0))$ to shift the activation threshold from 0 to v_0.

Another common choice for the activation function is the *hyperbolic tangent*,

$$tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{7}$$

Figure 4.3 shows the tanh function. A nice property of tanh function is $\frac{\partial tanh(x)}{\partial x} sech^2(x)$. In addition, the function has a nice Taylor series expansion and can be written as the fraction given below:

$$tanh(x) = x - \frac{1}{3}x^2 + \frac{2}{15}x^5 + \frac{17}{315}x^7 + \frac{62}{2835}x^9 - \dots \tag{8}$$

$$\frac{x}{1 + \frac{x^2}{3 + \frac{x^2}{5 + \dots}}} \tag{9}$$

Figure 4.2 The activation function $h(v)$ of Equation (4.4): the solid line is $h(v)$, the dashed line is $h(O.5v)$, and the dotted line is $h(5v)$.

8.1 Estimation or Training of Neural Networks

Application of neural networks often divides the data into training and forecasting subsamples. The parameters (or weights) are estimated using data in the training subsample. The fitting is typically carried out by the back propagation method, which is a gradient descent procedure. For a given $p - m - k$ network, the weights consist of $\{\alpha_{0i}\alpha_i | i = 1, \dots, m\}$ with $m(p + 1)$ elements and $\{\beta_{0j}, \beta_j | j = 1, \dots, k\}$ with $k(m + 1)$ elements. For ease in notation, we incorporate α_{0i} and β_{0j}; into α_i; and β_j;, respectively, and define $h_t = (h_{0t}, h_{1t}, \dots, h_{mt})'$, where $h_{0t} = 1$ and $h_{it} = h(\alpha_i' x_t)$ for $i > 0$, with the understanding that x, includes 1 as its first element.

Let θ be the collection of all weights. For continuous output variables, the objective function of model fitting is the sum of squared errors

$$\ell(\theta) = \sum_{j=1}^{k} \sum_{t=1}^{N} \left[y_{jt} - Y_j(x_t)\right]^2 \tag{10}$$

Figure 4.3. The hyperbolic tangent function.

where y_{jt} is the tth observation of the jth output variable, x_t denotes the tth observation of the input vector (including 1), and N denotes the sample size of the training subsample. For classification problems, we use the cross-entropy (or deviance) as the objective function,

$$\ell(\theta) = -\sum_{j=1}^{k} \sum_{t=1}^{N} y_{jt} \log Y_j(x_t) \tag{11}$$

and the classification rule is $argmax_j Y_j(x_t)$.

In the following, we use the objective function in Equation (4.7) to describe the back propagation method. Rewrite the objective function as

$$\ell(\theta)=\sum_{t=1}^{N}\ell_t,\ \ \text{with}\ \ \ell_t=\sum_{j=1}^{k}\left[y_{jt}-Y_j(x_t)\right]^2 \tag{12}$$

Taking partial derivatives and using the chain rule, we have

$$\frac{\partial\ell_t}{\partial\beta_{ij}}=-2\left[y_{jt}-Y_j(x_t)\right]g_j'\left(\beta_j'h_t\right)\beta_{vj}h'\left(\alpha_v'x_t\right)x_{it} \tag{13}$$

From Equations (4.9) and (4.10), a gradient decent update from the uth to the $(u+1)$th iteration assumes the form

$$\beta_{ij}^{u+1}=\beta_{ij}^{u}-\delta_u\sum_{t=1}^{N}\frac{\partial\ell_t}{\partial\beta_{ij}^{(u)}} \tag{14}$$

$$\alpha_{iv}^{u+1}=\alpha_{iv}^{u}-\delta_u\sum_{t=1}^{N}\frac{\partial\ell_t}{\partial\alpha_{iv}^{(u)}} \tag{15}$$

where δ_u is the learning rate, which is usually taken to be a constant and, if necessary, can also be optimized by a line search that minimizes the error function at each update. On the other hand, δ_u, should approach zero as $u\to\infty$ for online learning.

Rewrite Equations (4.9) and (4.10) as

$$\frac{\partial\ell_t}{\partial\beta_{ij}}=d_{jt}h_{it} \tag{16}$$

$$\frac{\partial\ell_t}{\partial\alpha_{iv}}=s_{vt}x_{it} \tag{17}$$

The quantities d_{jt} and s_{vt} can be regarded as errors from the current model at the output and hidden layer nodes, respectively. From their definitions, these errors satisfy

$$s_{vt}=h'\left(\alpha_v'x_t\right)\sum_{j=1}^{k}\beta_{vj}d_{jt} \tag{18}$$

which are the *back-propagation* equations from which the updates in Equations (4.11) and (4.12) can be carried out by a two-pass algorithm. Specifically, in the forward pass, the current weights are given and the predicted values $\hat{Y}_j(xt)$ are computed from the model in Equations (4.1)-(4.3). In the backward pass, the errors d_{jt} are computed and then back-propagated via Equation (4.15) to give the errors s_{vt}. The two sets of errors are then used to compute the gradients for the updates in Equations (4.11) and (4.12). The back-propagation is relatively simple, but it may converge slowly in an application. In addition, some cares must be exercised in training neural networks. Readers are referred to Hastie et al. (2001, Chapter 11) for further discussions and examples of application.

The gradient decent updates in Equations (4.11) and (4.12) can be modified by adding a random disturbance term to overcome some difficulties or to improve efficiency in network training.

4.1.2 An Example

To demonstrate the application of neural networks, we consider the US weekly crude oil prices from January 3, 1986 to August 28, 2015 for 1548 observations. The data are available from the FRED (Federal Reserve Bank of St. Louis) and are the West Texas Intermediate prices, Cushing, Oklahoma. Let x, denotes the weekly crude oil price at time index t. The time plot shows that x_t, has an upward trend during the sample period so we employ the change series $y_t = x_t - x_{t-1}$ with sample size $T = 1547$. Figure 4.4 shows the time plot of y_t, with the vertical line denoting the separation between modeling and forecasting subsamples.

Figure 4.4 Time plot of the change series of weekly US crude oil prices, West Texas Intermediate, Cushing, Oklahoma, from January 1986 to August 2015. The vertical line separates the modeling and forecasting subsamples.

Table 4.1 Out-of-sample predictions of various models for the change series of US weekly crude oil prices from January 1986 to August 2015. The last 200 observations are used as the forecasting subsample. RMSE, root mean squared error; MAE, mean absolute errors; AR(O), sample mean of the modeling subsample used as prediction.

The forecasting subsample consists of the last 200 observations. From the plot, it is clear that the variability of the price changes y, increased markedly after 2008.

For comparison, we use the root mean squared forecast errors (RMSE) and mean absolute forecast errors (MAE) to measure the performance of the model used in out-of-sample prediction. In addition, we use a linear AR(O) model (e.g., a white noise model) as a benchmark, and use two additional linear AR models. An AR(8) or AR(13) model fits the training data well. For the neural networks, we use the R package nnet in this example and employ $y_{t-1}, \ldots, Y_{t-13}$ as the input variables. Different numbers of nodes in the hidden layer are tried. Furthermore, we trained a given neural network multiple times to better understand its performance. Table 4.1 summarizes the out-of-sample performance of various models.

From the table, the neural networks, in this particular case, produce similar forecasts as the linear AR models. As expected, the models used fare better than the benchmark AR(∞).

R demonstration: Some commands used. The R command NNsetting of the NTS package is used to set up the input matrix and the output variable in both the training and forecasting subsamples for a time series.

```
da <- read.table('w-coilwti.txt', header=T)
yt <- diff(da[,4])
library(NTS)
m1 <- NNsetting(yt, nfore=200, lags=c(1:13) )
names (m1)
X <- m1$X
y <- m1$y
predX <- m1$predX
predY <- m1$predY
m2 <- lm(y~ -1+X)
yhat <- predY <- predX %*% matrix(m2$coefficients,13,1)
er1 <- yhat-predy
mean(er1^2)
mean(abs(er1))
library(nnet)
m3 <- nnet(X,y, size=1, skip=T, linout=T, maxit=1000)
pm3 <- predict(m3, newdata=predX)
er2 <- pm3-predY
mean(er2^2)
mean(abs(er2))
```

Davidson, Russell, and James G. MacKinnon. 2004. *Econometric Theory and Methods*. Oxford: Oxford University Press.

Enders, Walter. 2010. *Applied Econometric Time Series*. 3rd ed. New York: John Wiley & Sons.

Engel, Charles. 1994. “Can the Markov Switching Model Forecast Exchange Rates?” *Journal of International Economics* 36 (1-2): 151–65.

Engel, Charles, and James D Hamilton. 1990. “Long Swings in the Dollar: Are They in the Data and Do Markets Know It?” *American Economic Review* 80 (4): 689–713.

Filardo, Andrew J. 1994. “Business-Cycle Phases and Their Transitional Dynamics.” *Journal of Business & Economic Statistics* 12 (3): 299–308.

Franses, Philip Hans, and Dick van Dijk. 2000. *Non-Linear Time Series Models in Empirical Finance*. Cambridge: Cambridge University Press.

Granger, Clive, and Timo Teräsvirta. 1993. *Modelling Non-Linear Economic Relationships*. Oxford University Press.

Greene, William H. 2003. *Econometric Analysis*. New York: Prentice Hall.

Hamilton, James D. 1989. “A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle.” *Econometrica* 57: 357–84.

———. 1994. *Time Series Analysis*. Princeton: Princeton University Press.

Hayashi, Fumio. 2000. *Econometrics*. Princeton: Princeton University Press.

Judge, G. G., W. E. Griffiths, R. Carter Hill, H. Lutkepohl, and T. Lee. 1985. *The Theory and Practice of Econometrics*. Second. New York: Wiley.

Moolman, Elna. 2004. “A Markov Switching Regime Model of the South African Business Cycle.” *Economic Modelling* 21 (4): 631–46.

Shen, Chung-Hua, and David R. Hakes. 1995. "Monetary Policy as a Decision-Making Hierarchy: The Case of Taiwan." *Journal of Macroeconomics* 17 (2): 357–68.

---

1. See, Franses and Dijk (2000).↩
2. See, Enders (2010).↩
3. See, Davidson and MacKinnon (2004).↩
4. See, Greene (2003).↩
5. See, Franses and Dijk (2000).↩
6. This property is derived from the theory of ergodic Markov chains, where the eventual distribution of states is independent of the initial state. This property will not be discussed as part of this section of the course, but interested readers are referred to Hamilton (1994).↩
7. See also, Hamilton (1994).↩