Financial Econometrics
Lecture 9: Markov switching models and Artificial Neural Network models

Richard G. Pierse

1 Introduction

In this lecture we look at a very popular switching regimes model, due to Hamilton (1989) in which the regime in unobserved but is assumed to follow a simple stochastic process. We also look at *Artificial Neural Network* (*ANN*) models which were pioneered in economics by Hal White (1992) and which can be regarded as smooth transition models with many regimes, although because the models are semi-parametric only, the nature of the different regimes is unclear.

1.1 The *Markov Switching Regime* model

Hamilton (1989) proposed a switching regime model in which the regime or state is unobserved and is determined by a stochastic process. In the simplest case, the model is given by

$$y_t = \begin{cases} \alpha_{01} + \alpha_{11} y_{t-1} + u_t, & s_t = 1 \\ \alpha_{02} + \alpha_{12} y_{t-1} + u_t, & s_t = 2 \end{cases}.$$

where s_t represents the state or regime and the disturbance u_t is assumed to be independently, identically distributed $u_t \sim iid(0, \sigma^2)$. The state s_t is unobserved and is assumed to follow a first order *Markov process* so that

$$\Pr(s_t = i | s_{t-1} = j) = \pi_{ij}, \quad i, j = 1, 2. \tag{1.1}$$

The value π_{ij} is known as the *transition probability* of moving to state i from state j and is assumed to be independent of time.

The transition probabilities can be represented in a 2×2 matrix

$$\mathbf{\Pi} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}.$$

Note that the transition probabilities must satisfy the condition that $\sum_i \pi_{ij} = 1$, $\forall j$. This implies that the columns of $\mathbf{\Pi}$ must all sum to one or

$$\boldsymbol{\iota}'\mathbf{\Pi} = \boldsymbol{\iota}'$$

where $\boldsymbol{\iota}$ is a 2×1 vector of ones, so that the transition matrix $\mathbf{\Pi}$ is *singular*.

The probability of being in a particular state at a point in time, can be represented by the 2×1 vector $\Pr(\mathbf{s}_t)$

$$\begin{aligned} \Pr(\mathbf{s}_t) &= \mathbf{\Pi} \Pr(\mathbf{s}_{t-1}) \\ &= \mathbf{\Pi}^k \Pr(\mathbf{s}_{t-k}) \end{aligned}$$

The log-likelihood function of the observations y_t , $t = 1, \cdots, T$, is given by

$$L = \sum_{t=1}^{T} \log f(y_t|y_{t-1})$$

where

$$f(y_t|y_{t-1}) = \sum_{i=1}^{2} f(y_t|s_t = i, y_{t-1}) \Pr(s_t = i|y_{t-1}) \tag{1.2}$$

and the conditional distribution of y_t given knowledge of the state s_t has the standard normal density

$$f(y_t|y_{t-1}, s_t = i) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2}(y_t - \alpha_{0i} + \alpha_{1i}y_{t-1})^2\} \ .$$

Note that, by definition, the *joint distribution* of y_t and s_t is

$$f(y_t, s_t = i|y_{t-1}, s_{t-1}) = f(y_t|y_{t-1}, s_t = i) \Pr(s_t = i|s_{t-1})$$

and the marginal distribution of y_t is

$$f(y_t|y_{t-1}, s_{t-1}) = \sum_{i=1}^{2} f(y_t, s_t = i|y_{t-1}, s_{t-1}) \, .$$

Therefore the probability that $s_t = i$, given information at time t, is simply the ratio of the two:

$$\begin{aligned} \Pr(s_t &= i|y_t) = \frac{f(y_t, s_t = i|y_{t-1}, s_{t-1})}{f(y_t|y_{t-1}, s_{t-1})} \\ &= \frac{f(y_t|s_t = i, y_{t-1}) \Pr(s_t = i|s_{t-1})}{\sum_{i=1}^{2} f(y_t|s_t = i, y_{t-1}) \Pr(s_t = i|s_{t-1})} \, . \end{aligned} \tag{1.3}$$

Once the model parameters have been estimated by maximum likelihood, it is possible to derive an estimate of the probablility of being in regime i in time period t, based on all the observations in the sample, y_t, $t = 1, \cdots, T$. This is the so-called *smoothed estimator*

$$\Pr(s_t = i | y_1, y_2, \cdots, y_T).$$

The Markov switching model (1.1) can be generalised to deal with cases with more than two regimes or where the disturbance variance is state specific. However, the estimation of a Markov switching model proceeds on the assumption that the number of regimes is known, so that it is not possible to test the number of regimes within the model.

2 Markov switching models of volatility

Various versions of a *Markov-Switching GARCH* model have been developed by Kim (1993) and Hamilton and Susmel (1994) among others. A general version is defined by

$$\sigma_t^2 = \begin{cases} \alpha_{01} + \alpha_{11} u_{t-1}^2 + \beta_{11} \sigma_{t-1}^2, & s_t = 1 \\ \alpha_{02} + \alpha_{12} u_{t-1}^2 + \beta_{12} \sigma_{t-1}^2, & s_t = 2 \end{cases}. \tag{2.1}$$

where the state s_t is unobserved and is assumed to follow the first order *Markov process* defined by (1.1).

3 Artificial Neural Network models

Another important class of nonlinear models that are becoming increasingly popular in the finance literature are *Artificial Neural Network* (*ANN*) models. They were first developed in neurophysiology as ways of representing the operation of the human brain but have subsequently been applied to many areas in economics and finance, in particular by White (1992). The main reason for the popularity of *ANNs* is their ability to approximate, arbitrarily closely, any nonlinear function. In empirical work, they have been used to forecast stock prices, exchange rates and in option pricing.

The basic *ANN* we consider is the *multilayer perceptron (MLP) model*

$$y_t = \gamma_0 + \sum_{j=1}^{q} \gamma_j \phi(\mathbf{x}_t' \boldsymbol{\beta}_j) + u_t \tag{3.1}$$

where $\phi()$ is a squashing function called an *activation function* that varies smoothly between between 0 and 1. Any appropriate function can be used as $\phi()$ but the

most common choice is the *logistic function*

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

The terminology used in *ANNs* derives from neurology. The k variables $\mathbf{x}_t$ are known as the *input layer* and the single variable y_t is the *output layer*. The q *activation functions* $\phi()$ represent a *hidden layer* of q units or *nodes* between the input and output layer that transforms the inputs into values lying between zero and one. The activation functions are analogous to neurons in a brain. The coefficients γ_j are known as *synaptic weights*. In the model (3.1), there is only one hidden layer and so the model is known as a *single hidden layer feedforward network* model, where the qualification *feedforward* refers to the fact that information flows only in one direction, from inputs to outputs. Models allowing *feedback* are possible but are rarely used in financial applications.

The number of hidden layers can be increased. For example

$$y_t = \gamma_{00} + \sum_{i=1}^{p} \gamma_{0i}\phi(\sum_{j=1}^{q} \gamma_{ij}\phi(\mathbf{x}_t^{'}\boldsymbol{\beta}_j)) + u_t \tag{3.2}$$

represents a *dual hidden layer feedforward network* model where there are q nodes in the first hidden layer and p in the second.

3.1 Estimating ANNs

The *ANN* model (3.1) can be estimated by minimising the sum of squared residuals

$$\min_{\boldsymbol{\theta}} s = \sum_{t=1}^{T} \widehat{u}_t^2 = \sum_{t=1}^{T} \{y_t - \gamma_0 - \sum_{j=1}^{q} \gamma_j \phi(\mathbf{x}_t^{'}\boldsymbol{\beta}_j)\}^2 \tag{3.3}$$

with respect to the $(k+1)q+1$ unknown parameters $\boldsymbol{\theta} = \{\gamma_0, \gamma_1, \cdots, \gamma_q, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_q\}$. This is a form of *nonlinear least squares* (*NLS*) estimation. One problem with estimation of nonlinear models like this is that there may be several *local minima* so that the numerical minimisation procedure might converge to one of these local minima rather than the *global minimum*. One check is to try starting the minimisation from different initial values for $\boldsymbol{\theta}$ to see whether it converges to the same point.

Alternatively, the model parameters can be estimated *recursively*. Making use of the notational shorthand,

$$f(\mathbf{x}_t; \boldsymbol{\theta}) = \gamma_0 + \sum_{j=1}^{q} \gamma_j \phi(\mathbf{x}_t^{'}\boldsymbol{\beta}_j),$$

the recursive estimate can be written as

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} + \lambda_t \frac{\partial f(\mathbf{x}_t; \widehat{\boldsymbol{\theta}}_{t-1})}{\partial \widehat{\boldsymbol{\theta}}_{t-1}} (y_t - f(\mathbf{x}_t; \widehat{\boldsymbol{\theta}}_{t-1})) \tag{3.4}$$

where $\widehat{\boldsymbol{\theta}}_t$ denotes an estimate of $\boldsymbol{\theta}$ based on the first t observations The recursion is initialised with a set of starting values $\widehat{\boldsymbol{\theta}}_0$. This recursive way of estimating the model parameters is known as *learning* or *back-propagation* and λ_t is the learning rate, which is chosen either to be constant or to decline with time. White (1989b) provides a proof that, under certain conditions, this learning algorithm will either converge to a local minimum of (3.3) or $\widehat{\boldsymbol{\theta}}_t$ will tend to infinity.

One problem with the estimating (or '*training*' to use the special terminology) of *ANN* models is that of *over-fitting*. While this is a potential problem in any econometric model, it is especially acute here since *ANNs* are so good at approximating the behaviour of any nonlinear function. If an *ANN* is over-fitted, then it will be trained to explain some of the *noise* in the estimating data set as well as the *signal*, and this will tend to make it bad at prediction in new data sets on which it wasn't trained. It is very difficult to know *a priori* how complex an *ANN* model is needed (in terms of the number of nodes in each layer and the number of hidden layers) for a particular problem.

3.2 Interpreting ANN models

It is very difficult to give any interpretation to the synaptic weights γ in an *ANN* model, or to relate these to the form of nonlinearity captured by the model. This has led to *ANN* models being thought of as *black boxes*: they may be good at forecasting but are not susceptible to analysis because we can't look inside to see how they work.

It is interesting to compare the *ANN* model to the *STAR* model, since both make use of smooth transition logistic functions. For example, consider the equation

$$y_t = \gamma_0 + \gamma_1 \phi(\beta y_{t-1}) + u_t,$$

which is a special case of the *ANN* model (3.1) where the number of nodes in the hidden layer, q, is 1 and where the input variable vector, $\mathbf{x}_t$, simply consists of the first lag of y_t. This is related to the *STAR* model of last week's lecture which can be rewritten as

$$y_t = \omega_0 + \omega_1 \phi(\gamma y_{t-1} + \gamma c) + u_t$$

where

$$\omega_0 = \alpha_{01} + \alpha_{11} y_{t-1}$$

and

$$\omega_1 = \alpha_{02} - \alpha_{01} + (\alpha_{12} - \alpha_{11})y_{t-1}.$$

More generally, *single hidden layer ANN* models can be thought of as smooth transition models with many regimes. However, the interpretation of the regimes in an *ANN* model is unclear whereas it is explicit in a *STAR* model.

3.3 Testing linearity in an *ANN* model

White (1989a) proposed a test of linearity in the *ANN* model (3.1). Adding linear terms to the equation, it can be rewritten as

$$y_t = \mathbf{x}_t^{'}\boldsymbol{\beta}_0 + \sum_{j=1}^{q} \gamma_j \phi(\mathbf{x}_t^{'}\boldsymbol{\beta}_j) + u_t$$

and the null hypothesis of linearity corresponds to the q restrictions that $\gamma_1 = \gamma_2 = \cdots = \gamma_q = 0$. The alternative hypothesis is that the model is a *single hidden layer perceptron* model.

References

[1] Hamilton, J.D. (1989), 'A new approach to the economic analysis of nonstationary time series subject to changes in regime', *Econometrica*, 57, 357–384.

[2] Hamilton, J.D. and R. Susmel (1994), 'Autoregressive conditional heteroskedasticity and changes in regime', *Journal of Econometrics*, 64, 307–333.

[3] Kim, C-J. (1993), 'Unobserved-components time series models with Markov-switching heteroskedasticity: changes in regime and the link between inflation rates and inflation uncertainty', *Journal of Business and Economic Statistics*, 11, 341–349.

[4] White, H. (1989a), 'An additional hidden unit test for neglected nonlinearity in multilayer feedforward networks', *Proceedings of the International Joint Conference on Neural Networks, Washington D.C., Vol. 2*, SOS Printing, San Diego, CA.

[5] White, H. (1989b), 'Some asymptotic results for learning in single hidden-layer feedforward network models', *Journal of the American Statistical Association*, 84, 1003–1013.

[6] White, H. (1992), *Artificial Neural Networks: Approximation and Learning Theory*, Blackwell, Oxford, UK.