Quantitative Return Forecasting

Abstract

This page and its main links provide an introduction to quantitative return forecasting and explain the web service tools that the Nematrian website makes available to help with this activity. It is based on a section within Kemp (2005).

Many different techniques exist for trying to predict or forecast the future movements of investment markets. These range from purely judgemental to purely quantitative approaches and from ones that concentrate on individual securities to ones that are applied to entire markets. Quantitative return forecasting can in effect be thought of as a special case of time series analysis.

Traditional time series analysis often assumes that there is a linear relationship between the different variables of interest and that this function exhibits *time stationarity*. The analysis then in effect typically becomes akin to use of traditional linear regression techniques.

Unfortunately, such models can only describe a relatively small number of possible market dynamics, in effect just regular cyclicality and purely exponential growth or decay. Such techniques typically seem to work rather poorly for direct identification of profitable investment strategies. Investment markets do show cyclical behaviour, but the frequencies of the cycles are often far from regular. It is easy to postulate variables that ought to influence markets, but much more difficult to identify ones that seem to do so consistently whilst at the same time offering significant predictive power. Relationships that work well over some time periods often seem to work less well over others. Perhaps this is not too surprising. If successful forecasting techniques were easy to find then presumably market prices would have already reacted, reducing or eliminating their potential to add value in the future.

Better, therefore, are likely to be more sophisticated, quantitative return forecasting tools including some, like *locally linear regression* tools, that do not rely on time stationarity. These tools are implicitly more akin to how non-quantitative investment managers think and therefore may be expected to work more effectively in the real world. It is possible that *neural networks* could also help, although Nematrian is somewhat more sceptical about how effective such tools might be in practice for investment problems (because unless carefully designed they may *overfit* any available data).

Contents

1. Introduction

[ReturnForecasting1]

1.1 Many different techniques exist for trying to *predict* or *forecast* the future movements of investment markets. These range from purely judgemental to purely quantitative approaches and from ones that concentrate on individual stocks to ones that are applied to sectors or entire markets. In this set of pages on the Nematrian website we cover some of the more quantitative tools that have been devised for this purpose. Many very clever people have spent a lot of time devising quantitative ways of forecasting future investment returns, so in these pages cover only some of the many tools and techniques that might be used in practice.

1.2 Quantitative return forecasting can be thought of as a special type of time series analysis. Hence many of the time series analysis tools that are used in other contexts may also be applied to quantitative investment analysis. Time series analysis can in turn be split into two main types, both of which are typically analysed in a mathematical context using *regression* techniques. These are:

(a) Analysis of the interdependence of two or more variables *measured at the same time*, e.g. whether high inflation is associated with high asset returns. The assumption here is that there is some other exogenous way in which we can form an opinion on, say, how inflation will move in the future, and we then use this together exogenous view, together with an understanding of the interdependency of inflation and the asset return we want to forecast or predict to work out the most appropriate investment stance to adopt. The tools used are conceptually similar to those used for risk measurement, except that with risk measurement we are typically seeking to understand the spread of the distribution rather than its mean drift.

(b) Analysis of the interdependence of one or more variables *measured at different times*, usually with some intuitive justification proposed for the supposed interdependence being claimed from the analysis. Such links (if they can be found and if they persist) can be used directly to identify profitable investment strategies (as long as the excess returns available from their use are not swamped by transactions costs).

1.3 A simple example of a problem of the type described in 1.2(a) might involve postulating that there was some a linear relationship involving two time series, x_t and y_t (for $t = 1, \ldots, n$, where t is a suitable time index) of the form $y_t = a + bx_t + e_t$ where the e_t are random errors each with mean zero, and a and b are unknown constants. The same relationship can be written in vector form as $y = a + bx + e$ where $x = (x_1, \ldots, x_n)^T$ is a vector of n elements corresponding to each element of the time series etc. In such a problem the y_t are called the *dependent* variables and the x_t the *independent* variables, as in the postulated relationship the y_t depend on the x_t not vice-versa.

Such a problem is most commonly solved by use of regression techniques, as explained in many statistics textbooks. If the e_t are independent identically distributed normal random variables with the same variance (and same zero mean) then the maximum likelihood estimators of a and b are are the values that minimise the sum of the squared forecast error, i.e. $\sum\left(y_t - (a + bx_t)\right)^2$. These are also known as their *least squares estimators*. More generally, we might adopt other ways of estimating these variables including minimising, say, the mean absolute deviation, which involves minimising $\sum|y_t - (a + bx_t)|$.

1.4 To convert this simple example into one of the sort described in 1.2(b) we might incorporate a one-period time lag in the above relationship, i.e. we would assume that stocks, markets and/or factors driving them exhibit *autoregression*.

1.5 Typically, the mathematical framework involved can most easily be explained using vectors, see below. Mathematically we assume that there is some equation governing the behaviour of the system $y_t = f(y_{t-1}, y_{t-2}, \dots)$. The y_t might now in general be vector quantities rather than scalar quantities, some of whose elements might be unobserved *state* variables. However, the simplest examples have a single (observed) series in which later terms depend on former ones.

1.6 Traditional time series analysis generally assumes, at that f exhibits *time stationarity* (meaning it has the same functional form for each t). More advanced variants might include regime shifts or the like, in which the model of the world as characterised by f can vary in some well defined manner.

1.7 We shall see later that time stationary models can only describe a relatively small number of possible market dynamics (in effect just *regular* cyclicality and purely exponential growth or decay). This is probably why traditional linear time stationary regression techniques seem to be rather less effective than one might hope at directly identifying profitable investment strategies.

1.8 Investment markets do show cyclical behaviour, but the frequencies of the cycles are often far from regular. It is easy to postulate variables that ought to influence markets, but much more difficult to identify ones that seem to do so consistently whilst at the same time offering significant predictive power. Relationships that work well over some time periods often seem to work less well over others.

1.9 Perhaps this is not too surprising. If successful forecasting techniques were easy to find then presumably this would already be well known and market prices would have already reacted, reducing or eliminating the potential of such forecasting techniques to add value in the future. In this field, as in other aspects of active investment management, it is necessary to stay one step ahead of others!

2. Traditional time series analysis

[ReturnForecasting2]

2.1 Consider first a situation where we only have one time series where we are attempting to forecast future values from observed past values. For example, the time series followed by a given variable might be governed by the following relationship, where the value at time *t* of the variable is denoted by $y_t = cy_{t-1} + w_t$ where c is constant.

2.2 This is a *linear first order difference equation*. A *difference equation* is an expression relating a variable y_t to its previous values. The above equation is *first order* because only the first lag (y_{t-1}) appears on the right hand side of the equation. It is *linear* because it expresses y_t as a linear function of y_{t-1} and the *innovations* w_t. w_t are often treated as random variables, but we do not always need to do this.

2.3 Such a model of the world is also an *autoregressive* model, with a unit time lag and is therefore typically referred to as an $AR(1)$ model. It is also time stationary, since c is constant. Nearly all linear time series analysis assumes time invariance. We could however introduce secular changes by assuming one of the variables on which the time series is based is a dummy variable linked to time. An example commonly referred to in the quantitative investment literature is a dummy variable set equal to 1 in January but 0 otherwise, to identify whether there is any 'January' effect.

2.4 If we know the value y_0 at $t = 0$ then we find using *recursive substitution* that $y_t = c^t y_0 + \sum_{j=1}^{t} c^{t-j} w_j$. We can also determine the effect of each individual w_t on, say, y_{t+j}, the value of y that is j time periods further into the future value than y_t. This is sometimes called the *dynamic multiplier* $\frac{\partial y_{t+j}}{\partial w_t} = c^j$. If $|c| < 1$ then such a system is *stable*, in the sense that the consequences of a given change in w_t will eventually die out. It is *unstable* if $|c| > 1$. An interesting possibility is the borderline case where $c = 1$, when the output variable y_{t+j} is the sum of its initial starting value and historical inputs.

2.5 We can generalise the above dynamic system to be a linear *p*'th order difference equation by making it depend on the first p lags along with the current value of the innovation (input value) w_t, i.e. $y_t = c_1 y_{t-1} + c_2 y_{t-2} + \cdots + c_p y_{t-p} + w_t$. This can be rewritten in vector/matrix form as a first order difference equation, but relating to a *vector*, if we define the vector as follows:

$$\mathbf{g}_t \equiv \begin{pmatrix} y_t \\ y_{t-1} \\ \cdots \\ y_{t-p+1} \end{pmatrix} = \begin{pmatrix} c_1 & c_2 & \cdots & c_p \\ 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & \cdots & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \cdots \\ y_{t-p} \end{pmatrix} + \begin{pmatrix} w_t \\ 0 \\ \cdots \\ 0 \end{pmatrix} \equiv \mathbf{F}.\mathbf{g}_{t-1} + \mathbf{v}_t$$

$$\Rightarrow \mathbf{g}_t = \mathbf{F}^t \mathbf{g}_0 + \sum_{j=1}^{t} \mathbf{F}^{t-j} \mathbf{v}_j$$

2.6 These sorts of dynamic systems have richer structures than simple scalar difference equations. For a *p*'th order equation we have: $y_{t+j} = \sum_{k=1}^{p} f_{1,k}^{(j+1)} y_{t-k} + \sum_{k=1}^{j} f_{1,1}^{(j-k)} w_{t+k}$ (if $f_{i,k}^{(j)}$ is the element in the *i*'th row and *k*'th column of $\mathbf{F}^j$). To analyse the characteristics of such a system in more detail, we first need to identify the *eigenvalues* of $\mathbf{F}$. These are the values of λ for which $|\mathbf{F} - \lambda \mathbf{I}| = 0$ where $\mathbf{I}$ is the identity matrix. They are the roots to the following equation:

$$\lambda^p - c_1 \lambda^{p-1} - c_2 \lambda^{p-2} - \cdots - c_{p-1} \lambda - c_p = 0$$

2.7 A *p*'th order equation such as this always has *p* roots, but some of these may be complex numbers rather than real ones, even if (as would be the case in practice for investment time series) all the c_j are real numbers. Complex roots correspond to cyclical (sinusoidal) behaviour. We can therefore have combinations of exponential decay, exponential growth and sinusoidal (perhaps damped or inflating) behaviour. For such a system to be stable we require all the eigenvalues λ to satisfy $|\lambda| < 1$, i.e. for their absolute values all to be less than unity.

2.8 Eigenvalues are closely associated with *principal components analysis*. All non-negative definite symmetric $n \times n$ matrices, $\mathbf{V}$, will have n non-negative eigenvalues $\lambda_1, \ldots, \lambda_n$ and associated eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ (the eigenvectors can sometimes be degenerate) that satisfy $\mathbf{V}\mathbf{x}_i = \lambda_i \mathbf{x}_i$. The eigenvalues can be the same in which case the eigenvectors can be degenerate. The eigenvectors are orthogonal (or can be chosen to be orthogonal if they are degenerate), so that any *n*-vector $\mathbf{p}$ can be written as $\mathbf{p} = p_1\mathbf{x}_1 + p_2\mathbf{x}_2 + \dots + p_n\mathbf{x}_n$.

2.9 The principal components are the eigenvectors of the relevant covariance matrix corresponding to the largest eigenvalues, since they explain the greatest amount of variance when averaged over all possible positions. This is because $\mathbf{p} = p_1\mathbf{x}_1 + p_2\mathbf{x}_2 + \dots + p_n\mathbf{x}_n$. There is no fundamental reason why all stocks should be given equal weight in this averaging process. Different weighting schemas result in different vectors being deemed 'principal'.

3. The spectrum and *z*-transform of a time series

[ReturnForecasting3]

3.1 An equivalent way of analysing a time series is via its *spectrum* since we can transform a time series into a frequency spectrum (and vice versa) using Fourier transforms. Take for example another sort of prototypical time series model, i.e. the *moving average* or *MA* model. This assumes that the output depends purely on an input series (without autoregressive components), i.e.:

$$y_t = \sum_{k=1}^{N} b_k w_{t-k+1}$$

3.2 There are three equivalent characterisations of a MA model:

(a) In the *time domain* - i.e. directly via the $b_1, \dots, b_N$.

(b) In the form of *autocorrelations*, i.e. $\rho_\tau = E\big((y_t - \mu)(y_{t-\tau} - \mu)\big)/\sigma^2$ (where $E(x)$ means the expected value of x and $\mu = E(y_t)$ and $\sigma^2 = E((y_t - \mu)^2)$. If the input to the system is a stochastic process with input values at different times being uncorrelated (i.e. $E(w_i w_j) = 0$ for $i \neq j$) then the autocorrelation coefficients become:

$$\rho_\tau = \begin{cases} \displaystyle\sum_{k=|\tau|+1}^{N} b_k b_{k-|\tau|} \Big/ \sum_{k=1}^{N} b_k^2 & \text{if } |\tau| \leq N \\ 0 & \text{if } |\tau| > N \end{cases}$$

(c) In the *frequency domain*. If the input to a MA model is an impulse then the spectrum of the output (i.e. the result of applying the discrete Fourier transform to the time series) is given by:

$$S = |1 + b_1 \exp(-2\pi i.f) + b_2 \exp(-2\pi i.2f) + \dots + b_N \exp(-2\pi i.Nf)|^2$$

3.3 It is possible to show that an *AR* model of the form described earlier has a power spectrum of the following form: $S = 1/|1 - c_1 \exp(-2\pi i.f) - c_2 \exp(-2\pi i.2f) - \dots - c_N \exp(-2\pi i.Nf)|^2$. The obvious next step in complexity is to have both AR and MA components in the same model, e.g. an $ARMA(M, N)$ model, of the following form:

$$y_t = \sum_{m=1}^{M} c_m y_{t-m} + \sum_{n=1}^{N} b_n w_{t-n}$$

3.4 The output of an $ARMA$ model is most easily understood in terms of the *z-transform*, which generalises the discrete Fourier transform to the complex plane, i.e.:

$$X(z) \equiv \sum_{t=-\infty}^{\infty} x_t z^t$$

3.5 On the unit circle in the complex plane the *z*-transform reduces to the discrete Fourier transform. Off the unit circle, it measures the rate of divergence or convergence of a series. Convolution of two series in the time domain corresponds to the multiplication of their *z*-transforms. Therefore the *z*-transform of the output of an $ARMA$ model is:

$$Y(z) = C(z)Y(z) + B(z)W(z) = \frac{B(z)}{1 - C(z)} W(z)$$

3.6 This has the form of an input *z*-transform $W(z)$ multiplied by a *transfer function* $T(z) = B(z)(1 - C(z))^{-1}$ unrelated to the input. The transfer function is zero at the zeros of the MA term, i.e. where $B(z) = 0$, and diverges to infinity, i.e. has poles (in a complex number sense), where $C(z) = 1$, unless these are cancelled by zeros in the numerator. The number of poles and zeros in this equation determines the number of *degrees of freedom* in the model. Since only a ratio appears there is no unique $ARMA$ model for any given system. In extreme cases, a finite-order AR model can always be expressed by an infinite-order MA model, and vice versa.

3.7 There is no fundamental reason to expect an arbitrary model to be able to be described in an $ARMA$ form. However, if we believe that a system is linear in nature then it is reasonable to attempt to approximate its true transfer function by a ratio of polynomials, i.e. as an $ARMA$ model. This is a problem in function approximation. It can be shown that a suitable sequence of ratios of polynomials (called *Padé approximants*) converges faster than a power series for an arbitrary function. But this still leaves unresolved the question of what the *order* of the model should be, i.e. what values of M and N to adopt. This is in part linked to how best to approximate the *z*-transform. There are several heuristic algorithms for finding the 'right' order, for example the Akaike Information Criterion, see e.g. Billah, Hyndman and Koehler (2003). These heuristic approaches usually rely very heavily on the model being linear and can also be sensitive to the assumptions adopted for the error terms.

3.8 This point is also related to the distinction between *in-sample* and *out-of-sample* analysis. By in-sample we mean an analysis carried out on a particular data set not worrying about the fact that later observations would not have been know about at earlier times in the analysis. If, as will always be the case in practice, the data series is finite then incorporating sufficient parameters in the model will always enable us to fit the data exactly (in much the same way that a sufficiently high order polynomial can always be made to fit exactly a fixed number of points on a curve).

3.9 What normally happens is that the researcher will choose one period of time to estimate the parameters characterising the model and will then test the model *out-of-sample* using data for a subsequent (but still historic) time period. Sometimes the parameters will be fixed at the end of the in-sample period.

3.10 Alternatively, if we have some a priori knowledge about the nature of the linear relationship then our best estimate at any point in time will be updated as more knowledge becomes available in a Bayesian fashion. Updating estimates of the linear parameters in this manner is usually called applying a *Kalman filter* to the process, a technique that is also used in general insurance claims reserving.

3.11 In derivative pricing there is a similar need to avoid look-forward bias, and this is achieved via the use of so called *adapted* series, i.e. random series where you do not know what future impact the randomness being assumed will have until you reach the relevant point in time when the randomness arises.

3.12 However, it is worth bearing in mind that even rigorous policing of in-sample and out-of-sample analysis does not avoid an implicit element of 'look-forward-ness' when carrying out back tests of how a particular quantitative return forecasting might perform in the future. This is because the forecasters can be thought of as having a range of possible models from which they might choose. They are unlikely to present results where out-of-sample behaviour is not as desired. Given a sufficiently large number of possible model types, it is always be possible to find one consistent with an in-sample analysis that also looks good in a subsequent out-of-sample analysis. By the time we do the analysis we actually know what happened in both periods.

4. Generalising linear regression techniques

[ReturnForecasting4]

4.1 Multivariate regression involves the dependent variables (the y described earlier) depending on several different independent variables simultaneously. It can be thought of as mathematically equivalent to univariate regression, except with everything expressed using vectors rather than scalars.

4.2 There are several ways in which we can generalise linear regression, including:

(a) *Multiple regression*, in which the dependent variables depend on several different independent variables simultaneously;

(b) *Heteroscedasticity*, in which we assume that the e_t have different (known) standard deviations. We then adjust the weightings assigned to each term in the sum, giving greater weight to the terms in which we have greater confidence;

(c) *Autoregression*, in which the dependent data series depends not just on other independent data sets, but also on prior values of itself;

(d) *Autoregressive heteroscedasticity*, in which the standard deviations of the e_t vary in some sort of autoregressive manner;

(e) *Generalised linear least squares regression*, in which we assume that the dependent variables are linear combinations of (linear) functions of the x_t. Least squares regression is merely a special case of this, consisting of a linear combination of two functions $f_1(x_t) \equiv 1$ and $f_2(x_t) \equiv x_t$;

(f) *Non-normal random terms*, where we no longer assume that the random terms are distributed as normal random variables. This is sometimes called *robust regression*. This may involve distributions where the maximum likelihood estimators minimise $\sum|y_t - (a + bx_t)|$ in which case the formulae for the estimators then involve medians rather than means. We can in principle estimate the form of the dependency by the process of *box counting*, which has close parallels with the mathematical concept of entropy, see e.g. Press *et al*. (2007) or Abarbanel *et al.* (1993).

4.3 In all of the above refinements, if we know the form of the error terms and heteroscedasticity then we can always transform the relationship back to a *generalised linear regression* framework by transforming the dependent variable to be linear in the independent variables. The noise element might in such circumstances need to be handled using copulas and the like.

4.4 It is thus rather important to realise that only certain sorts of time series can be handled successfully within a linear framework however complicated are the adjustments that we might apply as above. All such linear models are ultimately characterised by a spectrum (or to be more a precise *z*-transform) that in general involves merely rational polynomials. Thus the output of all such systems is still characterised by noise superimposed on combinations of exponential decay, exponential growth, and regular sinusoidal behaviour.

We can in principle identify the dynamics of such systems by identifying the eigenvalues and eigenvectors of the corresponding matrix equations. If noise does not overwhelm the system dynamics we should expect the spectrum/z-transform to have a small number of distinctive peaks corresponding to relevant zeros or poles applicable to the AR or MA elements. We can postulate that these correspond to the underlying dynamics of the time series.

4.5 Noise will result in the spreading out of the power spectrum around these peaks. The noise can be ‘removed’ by replacing the observed power spectrum with one that has sharp peaks, albeit not with perfect accuracy (since we won’t know exactly where the sharp peak should be positioned). For these sorts of time series problems, the degree of external noise present is in some sense linked to the degree of spreading of the power spectrum around its peaks.

4.6 However, the converse is not true. Merely because the power spectrum is broad (and without sharp peaks) does not mean that its broadband component is all due to external noise. Irregular behaviour can still appear in a perfectly deterministic framework, if the framework is *chaotic*.

5. Chaotic market behaviour

[ReturnForecasting5]

5.1 To achieve *chaotic behaviour* (at least chaotic as defined mathematically) we need to drop the assumption of time stationarity, in some shape or form. This does not mean that we need to drop time predictability. Instead it means that the equation governing the behaviour of the system $y_t = f(y_{t-1}, y_{t-2}, \dots)$ involves a non-linear function f.

5.2 This change can create quite radically different behaviour. Take for example the *logistic map* or *quadratic map*: $y_t = cy_{t-1}(1 - y_{t-1})$ where c is constant. This mapping can also be thought of as a special case of generalised least squares regression (but not generalised *linear* least squares regression), in the sense that we can find c by carrying out a suitable regression analysis where one of the function is a quadratic. In this equation y_t depends deterministically on y_{t-1} and c is a parameter that controls the qualitative behaviour of the system, ranging from $c = 0$ which generates a fixed point ($y_t = 0$) to $c = 4$ where each iteration in effect destroys one bit of information.

5.3 To understand the behaviour when $c = 4$, we note that if we know the value to within ε (ε small) at one iteration then we will only know the position within 2ε at the next iteration. This exponential increase in uncertainty or divergence of nearby trajectories is what is generally understood by the term *deterministic chaos*. This behaviour is quite different to that produced by traditional linear models. Any broadband component in the power spectrum output of a traditional linear model has to come from external noise. With non-linear systems such output can be purely deterministically driven (and therefore in some cases predictable). The above example also shows that the systems do not need to be complicated to generate chaotic behaviour.

5.4 The main advantages of such non-linear models are that many factors influencing market behaviour can be expected to do so in a non-linear fashion and the resultant behaviour matches observations, e.g. markets often seem to exhibit cyclical behaviour, but with the cycles having no set lengths, and markets are often relatively little affected by certain drivers in some circumstances, but affected much more by the same drivers in other circumstances.

5.5 The main disadvantages of non-linear models are:

(a) The mathematics is more complex;

(b) Modelling underlying market dynamics in this way will make the modelling process less efficient if the underlying dynamics are in fact linear in nature; and

(c) If markets are chaotic, then this typically places fundamental limits on the ability of any approach to predict more than a few time steps ahead.

5.6 The last point arises because chaotic behaviour is characterised by small disturbances being magnified over time in an exponential fashion (as per the quadratic map described above with $c = 4$), eventually swamping the predictive power of any model that can be built up. Of course, in these circumstances using linear approaches may be even less effective!

5.7 Indeed, there are purely deterministic non-linear models that are completely impossible to use for predictive purposes even one step ahead. Take for example a situation in which there is a hidden state variable developing according to the following formula $x_t = 2x_{t-1}(mod\ 1)$ but we can only observe y_t, the integer nearest to x_t. The action of the map is most easily understood by writing x_t in a binary fractional expansion, i.e. $x_1 = 0.d_1d_2d_3 \ldots = d_1/2 + d_2/2^2 + d_3/2^3 + \cdots$. Each iteration shifts every digit to the right, so $y_t = d_t$. Thus this system successively reveals each digit in turn. Without prior knowledge of the seeding value, the output will appear to be completely random, and the past values of y_t available at time t tell us nothing at all about values at later times!

6. Neural networks

[ReturnForecasting6]

6.1 Mathematicians first realised the fundamental limitations of traditional time series analysis two or three decades ago. This coincided with a time when computer scientists were particularly enthusiastic about the prospects of developing artificial intelligence. The combination led to the development of *neural networks*.

6.2 A neural network is a mathematical algorithm that takes a series of inputs and produces some output dependent on these inputs. The inputs cascade through a series of steps that are conceptually modelled on the apparent behaviour of neurons in the brain. Each step ('neuron') takes as its input signals one or more of the input feeds (and potentially one or more of the output signals generated by other steps), and generates an output signal that would normally involve a non-linear function of the inputs (e.g. a logistic function). Typically some of the steps are intermediate.

6.3 Essentially any function of the input data can be replicated by a sufficiently complicated neural network. So it is not enough merely to devise a single neural network. What you actually need to do is to create lots of potential alternative neural networks and then develop some *evolutionary* or *genetic* algorithm that is used to work out which is the best one to use for a particular problem. Or, more usually, you define a much narrower class of neural networks that are suitably parameterised (maybe even just one class, with a fixed number of neurons and predefined linkages between these neurons, but where the non-linear functions within each neuron are parameterised in a suitable fashion). You then *train* the neural network, by giving it some historic data, adopting a *training algorithm* that you hope will home in on an appropriate choice of parameters that are likely to work well when attempting to predict the future.

6.4 There was an initial flurry of interest within the financial community in neural networks, but this interest seemed over time to subside. It is not that the brain doesn't in some respects seem to work in the way that neural networks postulate. Rather, earlier computerised neural networks generally proved rather poor at the sorts of tasks they were being asked to perform in this space.

6.5 More recently, with the advent of 'Big Data' and more powerful computers, there seems to have been a resurgence of interest in the topic of 'machine learning' and artificial intelligence. We can expect this to percolate into the financial community, if some firms identify approaches that seem successful with investment orientated problems. However, there is no guarantee that this will be easy. As Ghahramani (2015) notes, machine learning involves uncertainty, i.e. there is no certainty that investment orientated problems are easily amenable to such techniques, although possibly there are ways of modelling this uncertainty using the probabilistic framework to machine learning and therefore using probabilistic approaches to work out which types of investment problems are most amenable to machine learning techniques. He writes:

> "*The key idea behind the probabilistic framework to machine learning is that learning can be thought of as inferring plausible models to explain observed data. A machine can use such models to make predictions about future data, and take decisions that are rational given these predictions. Uncertainty plays a fundamental part in all of this. Observed data can be consistent with many models, and therefore which model is appropriate, given the data, is uncertain. Similarly, predictions about future data and the future consequences of actions are uncertain. Probability theory provides a framework for modelling uncertainty.*"

7. Locally linear time series analysis

[ReturnForecasting7]

7.1 One possible reason why neural networks were originally found to be relatively poor at financial problems is that the effective signal to noise ratio involved in such problems may be much lower than for other types of problem where they have proved more successful. In other words there is so much random behaviour that can't be explained by the inputs that they struggle to make much sense of it.

7.2 But even if this is not the case, it seems to me that disillusionment with neural networks was almost inevitable. Mathematically, our forecasting problem involves attempting to predict the immediate future from some past history. For this to be successful we must implicitly believe that the past does offer *some* guide to the future. Otherwise the task is doomed to failure. If the whole of the past is uniformly relevant to predicting the immediate future then, as we have noted above, a suitable transformation of variables moves us back into the realm of traditional linear time series, which we might in this context call *globally linear time series analysis*. To get the sorts of broadband characteristics that real time series return forecasting problems seem to exhibit you must therefore be assuming that some parts of the past are a better guide for forecasting the immediate future than other parts of the past.

7.3 This perhaps explains growth in interest in models that include the possibility of regime shifts, e.g. threshold autoregressive (TAR) models or refinements. These assume that the world can be in one of two (or more) states, characterised by, say, $y_t = f_1(y_{t-1}, y_{t-2}, \dots)$, $y_t = f_2(y_{t-1}, y_{t-2}, \dots)$, ... and that there is some hidden variable indicating which of these two (or more) world states we are in at any given time. We then estimate for each observed time period which state we were most likely to have been in at that point in time, and we focus our estimation of the model applicable in these instances to information pertaining to these times rather than to the generality of past history.

7.4 More generally, in some sense what we are trying to do is to:

(a) Identify the relevance of a given element of the past to forecasting the immediate future, which we might quantify in the form of some mathematical measure of 'distance', where the 'distance' between a highly relevant element of past and the present is deemed to be small, whilst for a less relevant element the 'distance' is greater; and

(b) Carry out what is now (up to a suitable transform) a *locally-linear time series analysis* (only applicable to the current time), in which you give more weight to those elements of the past that are 'closer', in the sense of (a), to present circumstances, see e.g. Abarbanel *et al.* (1993) or Weigend & Gershenfeld (1993).

7.5 Such an approach is *locally linear* in the sense that it involves a linear time series analysis but only using data that is 'local' (i.e. deemed relevant in a forecasting sense) to current circumstances. It is also implicitly how non-quantitative investment managers think. One often hears them saying that conditions are (or are not) similar to "the bear market of 1973-1994", "the Russian Debt Crisis", "the Asian crisis" etc., the unwritten assumption being that what happened then is (or is not) some reasonable guide to what might happen now.

7.6 Such an approach also:

(a) Caters for any feature of investment markets that you think is truly applicable in all circumstances, since this is the special case where we deem the entire past to be 'local' to the present in terms of its relevance to forecasting the future.

(b) Seems to encompass as special cases any alternative threshold autogressive model, because these can merely be thought of as special ways of partitioning up how such distances might be characterised.

7.7 Such an approach thus provides a true generalisation of traditional time series analysis into the chaotic domain.

7.8 This approach also provides some clues as to why neural networks might run into problems. In such a conceptual framework, the neural network training process can be thought of as some (relatively complicated) way of estimating the underlying model dynamics. A danger is that we start off with an initial definition of the class of neural networks that is then sifted through for a 'good fit' that is hugely over-parameterised. The training process should reduce this over-parameterisation, but by how much? If we fortuitously choose a good set of possible neural network structures to sift through, or if our training of the network is fortuitously good, then the neural network should perform well, but what are the odds of this actually occurring?

7.9 Of course, it can be argued that a locally linear time series analysis approach also includes potential over-parameterisation in the sense that there is almost unlimited flexibility in how you might define 'distance' between different points in time. Indeed, perhaps the flexibility here is mathematically equivalent to the flexibility of structure contained within the neural network approach, since any neural network training approach can be reverse engineered to establish how much weight is being given to different pasts for each component of the training data. However, the flexibility inherent in choice of 'distances' is perhaps easier for humans to visualise and understand than other more abstract ways of weighting past data.

7.10 Maybe the neural networkers had it the wrong way round. Maybe the neural networks within our brains are evolution's way of approximating to the locally linear framework referred to above. Or maybe consciousness, that elusive God-given characteristic of humankind, will forever remain difficult to understand from a purely mechanical or mathematical perspective.

References

[ReturnForecastingRefs]

Abarbanel, H.D.I. and Brown, R., Sidorowich, J.J. and Tsimring, L.S. (1993). The analysis of observed chaotic data in physical systems. *Reviews of Modern Physics*, Vol 65, No 4.

Billah, M.B., Hyndman, R.J. & Koehler, A.B. (2003). Empirical information criteria for time series forecasting model selection. *Monash University, Australia, Department of Econometrics and Business Statistics*, Working Paper 2/2003, ISSN 1440-771X.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, Vol 521

Kemp, M.H.D. (2005), Risk Management in a Fair Valuation World. *British Actuarial Journal*, 11, No. 4, pp. 595-712

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press

Weigend, A.S. & Gershenfeld, N.A. (1993). Time series prediction: forecasting the future and understanding the past, *SFI Studies in the Sciences of Complexity*, Proc Vol. XV, Addison-Wesley, 1993