Chapter 12

Non-Linear Models and their Applications

The course, so far, has concentrated on linear causal models:

$$X_t = \mu + \sum_{j=1}^{\infty} \psi_j Z_{t-j} \qquad \{Z_t\} \sim WN(0, \sigma^2).$$

There are many situations where a linear model does not fit and we need to consider more general models. The most general model would be:

$$X_t = f(t : Z_t, Z_{t-1}, Z_{t-2}, \ldots)$$

where $\{Z_t\} \sim WN(0, \sigma^2)$ and for each $t \in \mathbb{Z}$, $f(t; .)$ is a (deterministic) function. X_t is value of the process at time t, which depends on t and the innovations up to time t.

Let $\mathcal{F}_t$ denote the *observable* information up to time t; that is, $\mathcal{F}_t = \{X_t, X_{t-1}.X_{t-2}, \ldots\}$. We will restrict attention to situations where the conditional mean and variance of X_t given $\mathcal{F}_{t-1}$ may be written as:

$$\mu_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}] = g(\mathcal{F}_{t-1}), \qquad \sigma_t^2 = \mathrm{Var}(X_t|\mathcal{F}_{t-1}) = h(\mathcal{F}_{t-1}).$$

and where the model can be written as:

$$X_t = g(\mathcal{F}_{t-1}) + \sqrt{h(\mathcal{F}_{t-1})}\epsilon_t$$

where $\epsilon_t = \frac{Z_t}{\sigma_t}$ is the standardised shock, or standardised innovation. That is, $\{\epsilon_t\} \sim WN(0, 1)$.

12.1 Bilinear Model

The *bilinear model* is a model of the form:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j Z_{t-j} + \sum_{i=1}^{m} \sum_{j=1}^{s} \beta_{ij} X_{t-i} Z_{t-j} + Z_t.$$

Here p, q, m, s are non-negative integers. It is so-called, because for fixed Z, it is linear in X and for fixed X it is linear in Z. This level of linearity helps to estalish properties of the model and makes parameter estimation possible. The model was introduced by Granger and Andersen in 1978.

Consider now the following special bilinear model:

$$X_t = \mu + \phi X_{t-2} + \beta X_{t-2} Z_{t-1} + Z_t.$$

Using the fact that $Z_{t-1} \perp Z_t$ and $\{Z_{t-1}, Z_t\} \perp X_{t-2}$, the conditional mean and variance for this model can be computed quite easily; for $\{Z_t\} \sim WN(0, \sigma^2)$,

$$\mathbb{E}[X_t|\mathcal{F}_{t-2}] = \mu + \phi X_{t-2}, \qquad \mathrm{Var}(X_t|\mathcal{F}_{t-2}) = \left(1 + \beta^2 X_{t-2}^2\right)\sigma^2$$

If the further assumption is made that $\{Z_t\} \sim IIDN(0, \sigma^2)$ (independent identically distributed *normal* variables), then parameter estimation can be carried out using a quasi-likelihood. Here

$$X_t|\mathcal{F}_{t-2} \sim N(\mu + \phi X_{t-2}, (1 + \beta^2 X_{t-2}^2)\sigma^2)$$

so *conditionally*, conditioned on two time steps back, the likelihood function is straightforward. Removing the conditioning to obtain the marginal (and hence the usual likelihood) is not an easy problem, but we can use the conditional likelihoods instead. The parameter vector is $\theta = (\mu, \phi, \beta, \sigma)^t$ and the quasi-log likelihood is the following.

$$\begin{cases} L_n(\theta) = \sum_{t=1}^n l_t(\theta) \\ l_t(\theta) = -\frac{1}{2}\left(\log\sigma^2 + \log(1 + \beta^2 x_{t-2}^2)\right) + \frac{(x_t - \mu - \phi x_{t-2})^2}{\sigma^2(1+\beta^2 x_{t-2}^2)} \end{cases}$$

The maximising $\widehat{\theta}$ can be obtained by standard algorithms. It turns out that, asymptotically at least, this gives the right answer; asymptotic efficiency and asymptotic normality may be shown for $\widehat{\theta}$, the estimator of $\theta = (\mu, \phi, \beta, \sigma)^t$.

Example 12.1.

Consider the monthly returns of the CRSP equal-weighted index from January 1926 - December 2008 for 996 observations. Denote the series by X. Firstly, the sample pacf shows significant partial autocorrelations at lags 1 and 3, suggesting an AR(3) model.

Then, the squared series of the *residuals* of the AR(3) suggest that the conditional heteroskedasticity depends on lags 1,3 and 8 of the residuals. The special bilinear model:

$$X_t = \mu + \phi_1 X_{t-1} + \phi_3 X_{t-3} + (1 + \beta_1 Z_{t-1} + \beta_3 Z_{t-3}) Z_t$$

fits the data quite well. □

12.2 Threshold Autoregressive (TAR) Model

In practise, there are several non-linear characteristics that we would like to model: asymmetry in declining and rising patterns of a process. The *TAR* model uses threshold space to improve linear approximation. Consider a simple 2-regime AR(1) model:

$$X_t = \begin{cases} -1.5X_{t-1} + Z_t & X_{t-1} < 0 \\ 0.5X_{t-1} + Z_t & X_{t-1} \geq 0. \end{cases}$$

Here the *threshold* variable is X_{t-1} and the delay is 1; the threshold is 0.

A time series X_t is said to follow a k-regime *self-exciting TAR* (SETAR) model with threshold variable X_{t-d} if it satisfies:

$$X_t = \phi_0^{(j)} + \phi_1^{(j)} X_{t-1} + \ldots + \phi_p^{(j)} X_{t-p} + Z_t \qquad \gamma_{j-1} \leq X_{t-d} < \gamma_j.$$

Example 12.2.

The US monthly employment rate, seasonally adjusted and measured in percentage from January 1948 to March 2009 for 735 observations seems to follow a TAR model. A plot of the data shows two characteristis: slow upward trend and rapid decay. The series is not reversible and may not be unit-root stationary. The TAR model

$$Y_t = \begin{cases} 0.083Y_{t-2} + 0.158Y_{t-3} + 0.118Y_{t-4} - 0.180Y_{t-12} + Z_{1t} & Y_{t-1} \leq 0.1 \\ 0.421Y_{t-2} + 0.239Y_{t-3} - 0.127Y_{t-12} + Z_{2t} & Y_{t-1} > 0.1 \end{cases}$$

fits the data. The number of data points in regimes 1 and 2 are: 460 and 262.

12.3 Smooth Transition AR (STAR) Model

A time series X_t follows a 2-regime STAR(p) model if it satisfies:

$$X_t = c_0 + \sum_{i=1}^{p} \phi_{0i} X_{t-i} + F\left(\frac{X_{t-d} - \Delta}{s}\right)\left(c_1 + \sum_{i=1}^{p} \phi_{1i} X_{t-i}\right) + Z_t.$$

Here d is the delay parameter, Δ and s are parameters representing location and scale of model transition and $F(.)$ is a smooth transition function. In practise, F is either logistic, exponential or a cumulative distribution function.

The advantage of STAR over TAR is that the conditional mean function is differentiable; the disadvantage is that the parameters Δ and s are hard to estimate.

For both AR processes for TAR and STAR, the zeroes of the AR polynomials have to be outside the unit ball.

12.4 Markov Switching Model

A time series X_t follows a MSA (Markov Switching Autoregressive) model if it satisfies:

$$X_t = \begin{cases} c_1 + \sum_{i=1}^{p} \phi_{1i} X_{t-i} + Z_t & S_t = 1 \\ c_2 + \sum_{i=1}^{p} \phi_{2i} X_{t-i} + Z_t & S_t = 2 \end{cases}$$

Here S_t is a Markov chain with state space $\{1, 2\}$ with transition probabilities defined by

$$P_{12} = p_1, \qquad P_{21} = p_2.$$

12.5 Nonparametric Models

The essence of nonparametric models is *smoothing*. Consider two time series variable X and Y related by

$$Y_t = m(X_t) + Z_t$$

where m is an arbitrary function and $\{Z_t\} \sim WN(0, \sigma^2)$. We would like to estimate the unknown function m from the data. The most common technique is *kernel regression*. A *kernel* is a function $K \geq 0$ satisfying $\int K(y)dy = 1$. A *bandwidth* h is included;

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right).$$

The function m is estimated by:

$$\widehat{m}(x) = \frac{\sum_{t=1}^{T} K_h(x - X_t) Y_t}{\sum_{t=1}^{T} K_h(x - X_t)}.$$

Derivation Suppose we have a joint density $f_{X,Y}$ for (X, Y). We estimate this density by smoothing the empirical density; $e(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i, y_i}(x, y)$ where δ denotes a dirac delta mass function. This is smoothed in the following way: for each data point (x_i, y_i), we replace $\delta_{x_i,y_i}(x, y)$ by $K_h(x - x_i)\widetilde{K}_h(y - y_i)$, so that

$$\widehat{f}_{X,Y}(x, y) = \frac{1}{n} \sum_{j=1}^{n} K_h(x - x_j)\widetilde{K}_h(y - y_j).$$

Marginalising over y gives:

$$\widehat{f}_X(x, y) = \frac{1}{n} \sum_{j=1}^{n} K_h(x - x_j).$$

Now,

$$m(x) = \mathbb{E}[Y|X = x] = \int y f_{Y|X}(y|x)dy = \int y \frac{f_{X,Y}(x, y)}{f_X(x)} dy$$

which we approximate by

$$\widehat{m}(x) = \frac{\int y\widehat{f}_{X,Y}(x,y)dy}{\widehat{f}_X(s)} = \frac{\frac{1}{n}\sum_{i=1}^n K_h(x-x_i)\int y\widetilde{K}_h(y-y_i)dy}{\frac{1}{n}\sum_{i=1}^n K_h(x-x_i)}.$$

Using the fact that $\widetilde{K}_h$ is *symmetric*, therefore:

$$\widehat{m}(x) = \frac{\sum_{j=1}^n K_h(x-x_j)y_j}{\sum_{j=1}^n K_h(x-x_j)}.$$

□

Choice of Kernel Theoretical and practical considerations lead to a several possible kernels. One popular choice is the *Gaussian* kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}$$

another is the Epanechinkov kernel:

$$\tilde{K}(x) = 0.75\left(1-x^2\right)I\left(|x|\leq 1\right).$$

If there is a large quantity of data, then h is taken small; for $h = 0$

$$\widehat{m}(x) = \frac{\sum Y_t \mathbf{1}(X_t = x)}{\sum \mathbf{1}(X_t = x)}.$$

As $h \to +\infty$, $\widehat{m}(x) \overset{h\to+\infty}{\longrightarrow} \overline{Y}$. Large h leads to oversmoothing. The bandwidth is usually selected via a MISE (mean integrated squared error) criterion: minimising

$$MISE = \mathbb{E}\left[\int_{-\infty}^{\infty}\left(\widehat{m}(x)-m(x)\right)^2 dx\right]$$

This can be computed as $\widehat{h}_{opt} = 1.06\sigma n^{-1/5}$ for the Gaussian kernel and $\widehat{h}_{opt} = 2.34\sigma n^{-1/5}$ for the Epachinkov kernel. s, the estimate is used in place of σ.

Another method for bandwidth selection is *leave-one-out cross validation.*

12.6 Neural Networks

A *neural network* consists of

- input layer
- hidden layers
- output layer

If x_i denotes the value of the input of the ith node, the jth node of the hidden layer is given by:

$$h_j = f_j\left(\alpha_{0j} + \sum_{i\to j} w_{ij}x_i\right).$$

The *activation function* f_j is usually taken as:

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

For the *output*,

$$o = f_o(\alpha_{0o} + \sum_{j\to o} w_{jo}h_j)$$

where f_o can be *linear* $f_o(z) = z$ or *Heaviside* $f_o(z) = \mathbf{1}_{(0,+\infty)}(z)$. A neuron with a Heaviside function is called a *threshold neuron*, with 1 denoting that the neuron fires its message. For example, the output of the 2-3-1 network is:

$$o = \alpha_{0o} + w_{10}h_1 + w_{2o}h_2 + w_{3o}h_3$$

for a linear activation and

$$o = \begin{cases} 1 & \alpha_{)o} + w_{10}h_1 + w_{2o}h_2 + w_{30}h_3 > 0 \\ 0 & \alpha_{0o} + w_{1o}h_1 + w_{2o}h_2 + w_{3o}h_3 \leq 0 \end{cases}$$

if $f_o(.)$ is a Heaviside function.

Combining the layers, the output of a feed-forward neural network can be written as:

$$o = f_o\left(\alpha_{0o} + \sum_{j\to o} w_{jo}f_j\left(\alpha_{0j} + \sum_{i\to j} w_{ij}x_i\right)\right).$$

If one also allows for direct connections from the input layer to the output layer, then the network becomes:

$$o = f_o\left(\alpha_{0o} + \sum_{i\to o}\alpha_{io}x_i + \sum_{j\to o} w_{jo}f_j\left(\alpha_{0j} + \sum_{i\to j} w_{ij}x_i\right)\right).$$

The first summation is summing over the input nodes.

Training and Forecasting The first step is to build the network, determining the number of nodes, the biases α_{0j} and α_{0o} and weights w_{ij}. The second step is inference, especially forecasting.

In time series applications, let $\{(r_t, \mathbf{x}_t) : t = 1, \ldots, n\}$ denote the series of training data, where $\mathbf{x}_t$ denotes the vector of inputs, while r_t denotes the series of interest (e.g. log returns of a given asset). Training the network amounts to choosing these parameters to minimise a fitting criterion, for example least squares:

$$S^2 = \sum_{t=1}^{n} (r_t - o_t)^2.$$

This is a non-linear problem and may be approached by iterative methods. A popular algorithm is Back Propogation (BP), which starts with the output layer and works backwards, using a gradient rule to modify the parameters.