7 Nonlinear Models

Non-linearities can enter a model in a variety of ways. Some common examples are:

- ARMA models with multiplicative terms:

$$, y_t = \phi y_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1} y_{t-1}, \quad \varepsilon_t \sim NID(0, \sigma^2)$$

- Nonlinear functional forms
- Non-Gaussian, EG disturbances follow a t-distribution
- Dynamics in scale rather than location (e.g. GARCH)
- Switching regimes

There are 2 broad classes of models, parameter-driven and observation-driven.
An observation driven model is set up in terms of conditional distribution for the t-th observation: $p(y_t|Y_{t-1}|\phi)$. THe likelihood function is immediately available.
A parameter driven model typically does not allow for a likelihood function, where we have some link function: $y_t = \mu e^{\beta_t} + \varepsilon_t, \quad \beta_t = \phi\beta_{t-1} + \eta_t$, here it is exponential.

7.1 Nonlinear modelling and white noise

7.1.1 Law of iterated expectations

Definition 7.1.1: Law of iterated expectations (LIE)

$$\mathbb{E}[y] = \mathbb{E}_x[\mathbb{E}[y|x]]$$

Proof.

$$\mathbb{E}_x[\mathbb{E}[y|x]] = \int \left[\int yp(y|x)dy \right] p(x)dx = \int \int yp(y,x)dydx = \int yp(y)dy = \mathbb{E}[y]$$

□

This is useful, since we can find a sequence of one step ahead expectations:

$$\mathbb{E}_{t-j}[g(y_t)] = \mathbb{E}_{t-j} \cdots \mathbb{E}_{t-1}[g(y_t)]$$

where the unconditional expectation is found by letting $j \to \infty$.

Exercise 7.1.1. Show

$$Var[y] = \mathbb{E}_x[Var[y|x]] + Var_x[\mathbb{E}[y|x]]$$

Solution:-

$$\begin{aligned} Var[y] &= \mathbb{E}[y^2] - \mathbb{E}[y]^2 \\ &= \mathbb{E}_x[\mathbb{E}[y^2|x]] - \mathbb{E}_x[\mathbb{E}[y|x]]^2 \\ &= \mathbb{E}_x[Var[y|x] + \mathbb{E}[y|x]^2] - \mathbb{E}_x[\mathbb{E}[y|x]]^2 \\ &= \mathbb{E}_x[Var[y|x]] + \mathbb{E}_x[\mathbb{E}[y|x]^2] - \mathbb{E}_x[\mathbb{E}[y|x]]^2 \\ &= \mathbb{E}_x[Var[y|x]] + Var[\mathbb{E}[y|x]] \end{aligned}$$

7.1.2 White noise

White noise is uncorrelated, i.e. $\mathbb{E}[y_t y_s] = 0, \quad t \neq s$ with constant variance (and zero mean).
Strict white noise is stronger, we require independence, not just uncorrelatedness[1].
Martingale Difference has a zero (or constant) conditional expectation:

$$\mathbb{E}_{t-1}[y_t] = 0$$

and thus is uncorrelated with any function of past observations:

$$\mathbb{E}[y_t f(Y_{t-1})|Y_{t-1}] = f(Y_{t-1})\mathbb{E}[y_t|Y_{t-1}] = 0$$

Example.

$$y_t = \varepsilon_t + \beta\varepsilon_{t-1}\varepsilon_{t-2} \quad \varepsilon_t \sim NID(0, \sigma^2)$$

The autocovariance at lag τ can be derived as:

$$\begin{aligned} \mathbb{E}(y_t y_{t-\tau}) &= \mathbb{E}(\varepsilon_t + \beta\varepsilon_{t-1}\varepsilon_{t-2})(\varepsilon_{t-\tau} + \beta\varepsilon_{t-\tau-1}\varepsilon_{t-\tau-2}) \\ &= \mathbb{E}(\varepsilon_t\varepsilon_{t-\tau}) + \beta\mathbb{E}(\varepsilon_t\varepsilon_{t-\tau-1}\varepsilon_{t-\tau-2}) + \beta\mathbb{E}(\varepsilon_{t-1}\varepsilon_{t-\tau}\varepsilon_{t-\tau-1}) + \beta^2\mathbb{E}(\varepsilon_{t-1}\varepsilon_{t-2}\varepsilon_{t-\tau}\varepsilon_{t-\tau-1}) \\ &= 0 \quad \text{if} \quad \tau \neq 0 \end{aligned}$$

Since all observations are uncorrelated the series is white noise, however the observations are not independent, the conditional mean is:

$$\mathbb{E}_{t-1}[y_t] = \mathbb{E}_{t-1}[\varepsilon_t] + \beta\mathbb{E}_{t-1}[\varepsilon_{t-1}\varepsilon_{t-2}] = \beta\varepsilon_{t-1}\varepsilon_{t-2}$$

so the series is not a martingale difference.

Example (ARCH).

$$\begin{aligned} y_t &= \sigma_{t|t-1}\varepsilon_t, \quad \varepsilon_t \sim NID(0,1) \\ \sigma^2_{t|t-1} &= \gamma + \alpha y^2_{t-1} \end{aligned}$$

This is a Martingale difference since

$$\mathbb{E}_{t-1}[y_t] = \sigma_{t|t-1}\mathbb{E}_{t-1}[\varepsilon_t] = 0$$

implying it is also white noise.

[1] These are the same with Gaussian noise, since the distribution is fully defined by the first 2 moments

7.1.3 Linearity and Prediction

When disturbances in an ARMA are IID, the MMSE predictor is the conditional mean. It is linear in the observations and disturbances.
Assuming instead that the disturbances are MDs with mean zero and constant variance, the MMSE predictor is again the conditional expectation by the LIE.
When disturbances are WN (not strict WN) the MMSE = MMSLE, however if disturbances are not independent the MMSE is not the MMSLE.

Example.

$$y_t = \varepsilon_t + \beta\varepsilon_{t-1}\varepsilon_{t-2}, \quad \varepsilon_t \sim NID(0, \sigma^2)$$

The MMSLE of future observations is zero, with MSE equal to the variance of future observations:

$$\begin{aligned} Var(y_t) &= \mathbb{E}[\varepsilon_t^2] + 2\beta\mathbb{E}[\varepsilon_t\varepsilon_{t-1}\varepsilon_{t-2}] + \beta^2\mathbb{E}[\varepsilon_{t-1}^2\varepsilon_{t-2}^2] \\ &= \sigma^2 + 0 + \beta^2\sigma^4 \end{aligned}$$

However the MMSE (the conditional mean) is:

$$\mathbb{E}[y_t|Y_{t-1}] = \beta\varepsilon_{t-1}\varepsilon_{t-2}$$

which has MSE:

$$\mathbb{E}[y_t - \beta\varepsilon_{t-1}\varepsilon_{t-2}]^2 = \mathbb{E}[\varepsilon_t]^2 = \sigma^2$$

We can use the LIE to compute multi-step predictions:

$$\mathbb{E}_T[y_{T+\ell}] = \begin{cases} \mathbb{E}_T[\beta\varepsilon_T\varepsilon_{T-1}] = \beta\varepsilon_T\varepsilon_{T-1} & \ell = 1 \\ \mathbb{E}_T[\beta\varepsilon_{T+1}\varepsilon_T] = 0 & \ell = 2 \\ 0 & \ell > 2 \end{cases}$$

7.2 Stationarity

Theorem 7.2.1 (Krengel's Theorem). If y_t is strictly stationary and ergodic (SE) then a continuous transformation $g(y_t, y_{t-1}, \dots)$ is also SE.

Weak stationarity of g doesn't follow from weak stationarity of y_t since the moments may not exist. E.g. if $y_t \sim t\nu$ then $g(y_t) = e^{y_t}$ has no finite moments.

Definition 7.2.1: Linear stochastic recurrence equation

$$y_{t+1} = x_t y_t + z_t$$

where x_t and z_t are strictly stationary and ergodic.

Theorem 7.2.2. The conditions:

1. $\mathbb{E}(\max(0, \ln|z_t|)) = \mathbb{E}(\ln^+|z_t|) < \infty$
2. $\mathbb{E}(\ln|x_t|) < 0$

are sufficient for the existence and uniqueness of a strictly stationary solution for y_t.

Condition 1 usually holds, it is the second condition that is important. It is known as the **contraction condition** and can be interpreted as saying that the x_t's are on average smaller than 1

(in absolute value). We can see this by applying Jensen's inequality:

$$\mathbb{E}(\ln|x_t|) \leq \ln \mathbb{E}(|x_t|) \Rightarrow \mathbb{E}(|x_t|) < 1$$

Example. If x_t is lognormal, $lnx_t \sim N(\mu, \sigma^2)$ the contraction condition is $\mathbb{E}(\ln x_t) = \mu < 0$, whereat $\mathbb{E}(x_t) = e^{\mu+\sigma^2/2}$, thus $\ln \mathbb{E}(x_t) = \mu + \sigma^2/2 > \mu$ is stronger than needed.

Exercise 7.2.1. What is the stationarity condition for the bilinear model?

$$y_t = \phi y_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1} y_{t-1} \tag{7.1}$$

Solution:-

We can write the model as:

$$y_t = (\phi + \beta \varepsilon_{t-1}) y_{t-1} + \varepsilon_t$$

which is a SRE with $x_t = \phi + \beta \varepsilon_{t-1}$ and $z_t = \varepsilon_t$. The contraction condition is:

$$\mathbb{E}(\ln|x_t|) = \mathbb{E}(\ln|\phi + \beta \varepsilon_{t-1}|) < 0$$

7.2.1 Asymptotic Stationarity

Consider a non-linear generalisation of the linear SRE above:

$$y_{t+1} = \varphi(y_t, \mathbf{z}_t, \psi)$$

where $\mathbf{z}_t$ is a vector of SE variables[2] and ψ is a vector of parameters.

Theorem 7.2.3 (Bougerol's Theorem). If

- There exists y_1 such that $\mathbb{E}(\ln^+|\varphi(y_1, \mathbf{z}_1|)) < \infty$
- $\mathbb{E}(\ln \sup_y |\frac{\partial \varphi(y, \mathbf{z})}{\partial y}|) < 0$

then for any starting value y_1, a series y_t converges exponentially and almost surely to a unique SE solution. In other words $|y_t(y_1, \psi) - y_t(\psi)| \overset{e.a.s.}{\to} 0$ as $t \to \infty$.

Intuitively this just means that it doesn't matter where we start, we will always converge to the same solution.

Example. Consider this cursed AR(1) model:

$$y_t = \phi \frac{e^{y_{t-1}} - 1}{e^{y_{t-1}} + 1} + \varepsilon_t \quad \varepsilon_t \sim NID(0, \sigma^2)$$

Clearly

$$-1 < \frac{e^{y_{t-1}} - 1}{e^{y_{t-1}} + 1} < 1$$

[2] When $\mathbf{z_t}$ is a vector of IID variables, the equation is known as a Markov system

so the first condition holds for any finite y_1. We now examine the second condition:

$$\begin{aligned}\left|\frac{\partial}{\partial y}\left(\phi\frac{e^y-1}{e^y+1}+\varepsilon_t\right)\right| &= \left|\phi\frac{(e^y+1)e^y-(e^y-1)e^y}{(e^y+1)^2}\right| \\ &= |\phi|\left|\frac{2e^y}{(e^y+1)^2}\right| \\ &= |\phi|\frac{2e^y}{(e^y+1)^2}\end{aligned}$$

To solve for the supremum we take the derivative and set it to zero:

$$\begin{aligned}0 &= \frac{d}{dy}\frac{2e^y}{(e^y+1)^2} \\ \Rightarrow 0 &= (e^y+1)^2e^y - 2e^y(e^y+1)e^y \\ \Rightarrow e^y+1 &= 2 \\ \Rightarrow y &= 0 \quad \text{with} \quad \frac{2e^y}{(e^y+1)^2} = \frac{1}{2}\end{aligned}$$

Thus:

$$\mathbb{E}(\ln \sup_y |\frac{\partial\varphi(y,\mathbf{z})}{\partial y}|) = \mathbb{E}(\ln \frac{|\phi|}{2}) = \mathbb{E}(\ln|\phi|) - \ln 2 < 0 \equiv |\phi| < 2$$

7.3 Distributions

Definition 7.3.1: Survival function

$$S(y) = P(Y > y) = 1 - F(y)$$

Example (Exponential).

$$F(y) = 1 - e^{-y/\theta} \Rightarrow S(y) = e^{-y/\theta}$$

Definition 7.3.2: Probability integral transform

The PIT of Y is the standard uniform, i.e. $F(Y) \sim U(0,1)$

$$f(F(Y)) = f(y)\frac{dy}{dF(y)} = 1$$

Thus we can generate any distribution by transforming a standard uniform.

Definition 7.3.3: t-distribution

$$f(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{(y-\mu)^2}{\nu\phi^2}\right)^{-\frac{\nu+1}{2}}$$

where μ is median and ϕ is scale.

The location-dispersion model is:

$$y_t = \mu + \psi\varepsilon_t$$

where ε_t has mean zero, and ψ is called the dispersion for y_t. For non-negative variables a location + scale model is needed:

$$y_t = \psi\varepsilon_t$$

where ε_t has mean one. When $y_t > 0$ taking logarithms gives:

$$\ln y_t = \ln \psi + \ln \varepsilon_t$$

so $\ln \psi$ is now a location parameter.

Definition 7.3.4: Gamma distribution

A gamma(ψ, γ) distribution has density:

$$f(y) = \psi^{-\gamma} \frac{y^{\gamma-1} e^{-y/\psi}}{\Gamma(\gamma)}$$

The chi-squared distribution is gamma$(2, \nu/2)$, setting $\gamma = 1$ gives an exponential distribution.

Definition 7.3.5: Log-logistic distribution

A log-logistic(ψ, γ) distribution has density:

$$f(y) = \frac{\nu \frac{y}{\psi}^{\nu-1}}{\psi \left(1 + \left(\frac{y}{\psi}\right)^{\nu}\right)^2}$$

Definition 7.3.6: Quantiles

The α-quantile of a distribution is the value y_α such that $F(y_\alpha) = \alpha$. The median is the 0.5-quantile.

Definition 7.3.7: Heavy tails

A distribution is said to be heavy tailed if

$$\lim_{y \to \infty} \exp(y/\alpha) S(y) = \infty \quad \forall \alpha > 0$$

Example (Exponential distribution). $S(y) = e^{-y/\alpha}$ so

$$\exp(y/\alpha) S(y) = e^{y/\alpha} e^{-y/\alpha} = 1$$

thus it is not heavy tailed.

Definition 7.3.8: Fat tails

A distribution is said to have fat tails if

$$S(y) = cL(y) y^{-\alpha} \quad \alpha > 0$$

where $L(y)$ is slowly varying, i.e. $\lim_{y \to \infty} L(\lambda y)/L(y) = 1$ for all $\lambda > 0$ and c is a non-negative constant.

Claim 7.3.1. Fat tailed $\Rightarrow$ heavy tailed, but not the reverse.

7.4 Nonlinear state space models

Parameter driven models may be nonlinear in the measurement equation, the transition equation or both. The basic model is:

$$y_t = f(\theta_t, \varepsilon_t|\varphi)$$
$$\theta_{t+1} = \psi(\theta_t, \eta_t|\varphi)$$

where θ is the signal, φ parameters and ε, η disturbances with specified distributions.

Example (AR1 dynamic equation)**.** Consider

$$y_t = \mu \exp(\beta_t) + \varepsilon_t \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$$
$$\beta_{t+1} = \phi\beta_t + \eta_t \quad \eta_t \sim NID(0, \sigma_\eta^2) \quad |\phi| < 1$$

Clearly the state equation follows an AR1 with parameter less than 1, it is SE. We can thus apply Krengel's theorem to show that y_t is also SE.

Example (Non-negativity)**.** When y_t is non-negative, any model must be non-linear. Consider the measurement equation below, with time varying mean μ_t:

$$y_t = \mu_t \varepsilon_t \quad 0 \leq y_t < \infty$$

where ε_t has a gamma distribution with mean 1. We can use an exponential link function to model the logarithm of mu_t to ensure mu_t remains positive:

$$\ln \mu_{t+1} = \delta + \phi \ln \mu_t + \alpha\eta_t$$

The restriction $|phi| < 1$ guarantees the stationarity of $\ln \mu_t$, and hence of y_t.

We can also consider conditionally Gaussian state space models, allowing the possibility of feedback from past observations to the system matrices that would otherwise be a linear SSM. We can derive the Kalman filter exactly as before and construct the likelihood function since the system matrixes are fixed at time $t-1$. It is not usually possible to predict more than one-step ahead however.

Example. An example of a conditionally Gaussian process is

$$y_t = \phi_t y_{t-1} + \varepsilon_t$$

$$\phi_{t+1} = \phi(1-\alpha) + \alpha\phi_t + \eta_t$$

where α is a fixed parameter. When ϕ_t is regarded as the state, we have a conditionally Gaussian SSM with $z = y_{t-1}$. When $|\alpha| < 1$ the model is SE.

7.5 Observation driven models

An observation driven model is set up to give a conditional distribution for each observation, that is:

$$p(y_t|Y_{t-1}|\varphi)$$

where Y_{t-1} denotes all past observations. We can then construct the likelihood in the usual way.

Example (Bilinear model).

$$y_t = \phi y_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1} y_{t-1}$$

We can compute the ML estimate by minimising the SSR. Setting up the model in SSF:

$$y_t = \mu_{t|t-1} + \varepsilon_t$$
$$\mu_{t+1|t} = \phi y_t + \beta \varepsilon_t y_t = (\phi + \beta \varepsilon_t)\mu_{t|t-1} + \phi \varepsilon_t + \beta \varepsilon_t^2$$

More general observation models can be written:

$$p(y_t | \alpha_{t|t-1}, Y_{t-1} | \varphi)$$

with the transition equation becoming a filtering equation

$$\alpha_{t+1|t} = g(\alpha_{t|t-1}, y_t, Y_{t-1} | \varphi)$$

The signal θ is the dynamic parameter in the conditional distribution. It depends on $\alpha_{t+1|t}$ so

$$\theta_{t|t-1} = h(\alpha_{t|t-1}, y_t, Y_{t-1} | \varphi)$$

and we can write

$$p(y_t | h(\alpha_{t|t-1}, Y_{t-1} | \varphi))$$

7.5.1 Moment-driven and score-driven models

Consider the linear Gaussian AR1N model. The finite sample Kalman filter begins at $t = 1$ taking account of the uncertainty with a diffuse prior. However we might consider a model based on the steady-state (constant Kalman gain), but initialised at $t = 1$. We write this as:

$$y_t = \mu + \mu_{t|t-1} + \nu_t$$

$$\mu_{t+1|t} = \phi \mu_{t|t-1} + \kappa \nu_t$$

where the innovations ν_t are NID and $\mu_{1|0}$ is fixed. We can write the filter as

$$\begin{aligned} \mu_{t+1|t} &= \phi \mu_{t|t-1} + \kappa \nu_t \\ &= \phi \mu_{t|t-1} + \kappa (y_t - \mu - \mu_{t|t-1}) \\ &= (\phi - \kappa)\mu_{t|t-1} + \kappa (y_t - \mu) \end{aligned}$$

Substituting repeatedly gives

$$\mu t + 1|t = \kappa \sum_{j=0}^{t-1} (\phi - \kappa)^j (y_{t-j} - \mu) + (\phi - \kappa)^t \mu_{1|0}$$

Although the filter is typically started with $\mu_{1|0} = 0$, as long as $|\phi - \kappa| < 1$ the filtered level will converge to the same value for any starting value.
In a score-driven model the dynamic equation for the above is driven by a variable that is proportional to the score of the conditional distribution, that is:

$$\mu_{t+1|t} = \phi \mu_{t|t-1} + \kappa \left(k \frac{\partial \ln p(y_t | \mu_{t|t-1})}{\partial \mu_{t|t-1}} \right)$$

The inverse of the information matrix is a common choice for k.

Example (Non-negativity - Gamma distribution). For non-negative variables we use a location-scale model. Suppose $p(y_t|\mu_{t|t-1}, \psi)$ is gamma with shape parameter γ. A filter for the conditional mean can be written as:

$$\mu_{t+1|t} = \delta + \beta\mu_{t|t-1} + \alpha y_t$$

We can write the likelihood function as

$$L = \prod_{t=1}^{T} \frac{1}{\Gamma(\gamma)} \left(\frac{\gamma}{\mu_{t|t-1}}\right)^{\gamma} y_t^{\gamma-1} e^{-\frac{\gamma y_t}{\mu_{t|t-1}}}$$

and thus the log-likelihood is:

$$\ln L = -T \ln \Gamma(\gamma) + T\gamma \ln \gamma - \gamma \sum_{t=1}^{T} \ln \mu_{t|t-1} + (\gamma - 1) \sum_{t=1}^{T} \ln y_t - \gamma \sum_{t=1}^{T} \frac{y_t}{\mu_{t|t-1}}$$

Differentiating with respect to $\mu_{t|t-1}$ gives the score:

$$\frac{\partial \ln L}{\partial \mu_{t|t-1}} = -\frac{\gamma}{\mu_{t|t-1}} + \frac{\gamma y_t}{\mu_{t|t-1}^2} = \frac{\gamma(y_t - \mu_{t|t-1})}{\mu_{t|t-1}^2}$$

HOW DOES THIS GIVE $I(\mu) = \frac{\gamma}{\mu_{t|t-1}^2}$?