1 AR(1) Models

1.1 Introduction

- Time series: Stochastic process in (discrete) time, sometimes (often) equally spaced.
- Scalar time series, continuous measurements - continuous state space.
- Simplest non-trivial model: AR(1) - AutoRegressive of order 1.
- Time series is “regressed” on itself - prior value predicts current value.

For $t = 0, 1, \ldots$, and in principle for theoretical development, $t = -1, -2, \ldots$,

$$x_t = \phi x_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, v).$$  \hfill (1)

- AR parameter $\phi$.
- Innovation (error, evolution error, stochastic input at time $t$) $\epsilon_t$.
- Innovations are independent, Gaussian (normal), zero-mean and constant variance, and independent of $x_{t-1}$ and all past $x_{t-j}$. Independence is often written as $\epsilon_t \perp \epsilon_s$ (for all $t, s$).
- Innovation is time $t$ random “shock,” added to “predicted” value for $x_t$ given by $\phi x_{t-1}$. The shock is “unpredictable” in the sense that $E(\epsilon_t) = 0$, and the innovations variance $v$ defines the scale of randomness injected into the evolution of the $x_t$ process at each time point.
- Look at models, structure, distribution theory assuming that the model parameters $\theta$ defined by $\theta = (\phi, v)$ are known. Model analysis in applications involves estimation of parameters and model assessment.
- Notation (not yet, perhaps, standard, but useful): $x_t \leftarrow AR(1|\theta)$, or $AR(1|(\phi, v))$.
- More general models could have different innovations variances at each time, or even non-normal distributions for the innovations.

AR(1) models are of major interest in their own right as simple stochastic process models for many time series applications, but also of very major importance as building blocks of more complex models representing real phenomena. The model class is also a nice setting to introduce core ideas of multivariate distribution theory linked to normal models, to develop initial ideas of simulation - of univariate and structure multivariate distributions, stochastic processes and then posterior distributions for parameters arising in model fitting and Bayesian inference. The AR(1) model class is an example of a class of Markov (Markovian) stochastic processes on a continuous (univariate) state space, which provides - among other things - key examples of Markov chains, and entree to the ideas and theory of Markov chains and of Markov Chain Monte Carlo (MCMC) simulation methods. As components of more complicated probability models, AR(1) models can become “hidden” (latent) processes, so providing examples of hidden Markov models (HMMs).
1.2 Structure and Distribution Theory in Stationary AR(1) Processes

1.2.1 Stationarity

Stationarity of the process means that the \(n\)-variate joint distribution of \(x_{s:s+n-1} = (x_s, x_{s+1}, \ldots, x_{s+n-1})'\) does not depend on \(s\), for any \(n \geq 1\). Weak stationarity refers to the mean and variance-covariance of the joint distribution, but in the case of linear, normal models those moments characterize the full joint distribution. In particular,

- \(n = 1\): each \(x_t\) has the same distribution,
- \(n = 2\): Each pair of values \(x_t, x_s\) has the same bivariate distribution,

and in the current context these are all normal distributions.

1.2.2 Conditional and Marginal Univariate Distributions

Critical to distinguish distributions via the conditioning elements. All distributions are implicitly conditioned on the specified parameter values, for now. Model equation (1) specifies, for all \(t\), the conditional distribution

\[ p(x_t|x_{t-1}) = N(\phi x_{t-1}, \sigma). \]  

(2)

The first-order Markovian property is that, conditional on \(x_{t-1}\), the distribution of \(x_t\) does not depend on previous values \(x_{t-j}\) for \(j > 1\). This is often written as \(x_t \perp \perp x_{t-j}, (j > 1)|x_{t-1}\), although a perhaps clearer notation is \((x_t|x_{t-1}) \perp \perp x_{t-j}, j > 1\).

Assuming weak stationarity alone, write \(m = E(x_t)\) and \(s = V(x_t)\) so that for all \(t\) we know the marginal distribution \(x_t \sim N(m, s)\). We can see that

- \(m = E(x_t) = E[E(x_t|x_{t-1})] = E(\phi x_{t-1}) = \phi m\), which can only hold if \(m = 0\) unless \(\phi = 1\), a very special nonstationary random walk case.
- Similarly, \(s = V(x_t) = E[V(x_t|x_{t-1})] + V[E(x_t|x_{t-1})] = v + \phi^2 s\) so that \(s = v/(1 - \phi^2)\). This can only make sense if \(|\phi| < 1\), characterising stationary AR(1) models.

1.2.3 Linear Process

The model is a linear time series model, or linear process. Iterate equation (1) to get

\[ x_t = \epsilon_t + \phi \epsilon_{t-1} + \ldots + \phi^k \epsilon_{t-k} + \ldots \]  

(3)

The process is linear - a linear function of current and past innovations, and a sum of independent stochastic elements that are weighted by the AR parameter. If \(|\phi| < 1\) the weight at lag \(k\) decays as \(k\) increases so that the current value of the \(x\) process is less and less dependent on the past innovations. An explosive (quite nonstationary) process results otherwise.

Linear processes can be non-Gaussian. This framework is Gaussian. Imagine an AR(1) model in which the innovations are independent but from some non-Gaussian distribution, such as a Student T or Cauchy, or other.
1.2.4 Backshift Operator

**Backshift operator notation and manipulation:** $Bx_t = x_{t-1}$ and $B^k x_t = x_{t-k}$ for all $k > 0$. The model then can be written as $(1 - \phi B)x_t = \epsilon_t$, and this becomes useful for formal manipulations. In algebraic equations the $B$ operator can be treated as if it were a number in $(0,1)$: $x_t = (1 - \phi B)^{-1} \epsilon_t$ and the expansion $(1 - \phi B)^{-1} = 1 + \phi B + \phi^2 B^2 + \ldots$ lead to the (stationary, lagged weights decaying with time) linear representation equation (3).

1.3 Autocorrelations and Full Joint Distributions

- Covariance at lag $k$: $\gamma(k) = C(x_t, x_{t\pm k})$
- $\gamma(k) = \phi^k$
- Correlation at lag $k$: $\rho(k) = \phi^k$
- Use linear representation, or iterated covariances to derive. Autocorrelation at first lag is the defining AR parameter $\phi$.
- Not well-defined in nonstationary processes, though some nonstationary processes do have definable conditional autocorrelations.

Any set of $x$ values has a joint normal distribution, a key example is that for (any) $n$ consecutive values, such as $x_{1:n} = (x_1, x_2, \ldots, x_n)'$, a column $n-$vector. We know the mean (0) and the variances and covariances, and the linear representation means that $x_{1:n}$ is a linear function of independent normal innovations so is multivariate normal (see multivariate normal theory reference material too). We write

$$x_{1:n} \sim N(0, \Sigma_n)$$  \hspace{1cm} (4)

where 0 is now the $n-$vector of zeros and $\Sigma_n$ is the variance (or variance-covariance matrix, a symmetric positive definite - SPD - $n \times n$ matrix), which is $\Sigma_n = s\Phi_n$ with correlation matrix

$$\Phi_n = \begin{pmatrix}
1 & \phi & \phi^2 & \ldots & \phi^{n-1} \\
\phi & 1 & \phi & \ldots & \phi^{n-2} \\
\phi^2 & \phi & 1 & \ldots & \phi^{n-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\phi^{n-1} & \phi^n & \phi^{n-2} & \ldots & 1
\end{pmatrix}. \hspace{1cm} (5)$$

Some general considerations related to existence of these joint distributions, stationarity and nonstationarity, and also structure of the stationary model:

- A nice exercise in linear algebra: Find $\Phi_n^{-1}$.
- What happens if $\phi = 0$?
- Many applications have $\phi > 0$, but some will involve “oscillatory” behaviour consistent with $\phi < 0$.
- What about the random walk limit at $\phi = 1$? The model is well-defined but nonstationary, and this is a very, very important, simple model. Look at $p(x_t|x_0)$ for $t > 0$.

1.4 Bivariate Distributions, Markov Transitions & Reversibility

$$p(x_{t-1}, x_t) = p(x_t|x_{t-1})p(x_{t-1})$$  \hspace{1cm} (6)

- $p(x_t|x_{t-1})$ is normal, linear regression from model equation (1).
- $p(x_{t-1})$ is (stationary) marginal $N(0, s)$. 

3
• Fill in density functions to show the bivariate density is log-quadratic, a normal density.

**Key identity for stationary processes:** We know that, always in any bivariate distribution,

\[ p(x_t) = \int p(x_t|x_{t-1})p(x_{t-1})dx_{t-1}. \]  

(7)

In this special framework of stationary (linear, normal) AR(1) models, the two univariate margins are the same distribution, \( N(0,s) \). The conditional distribution defined by the model is also named the transition distribution of the Markov process (or evolution distribution). Changing notation a little to highlight this very general, key representation, write \( g(u) \) for the density function of \( u \sim \mathcal{N}(0,s) \), and \( t(x|u) \) for the conditional normal density of \( (x|u) \sim \mathcal{N}(\phi u, v) \), then the key identity above is

\[ g(x) = \int t(x|u)g(u)du. \]  

(8)

The bivariate density is then, in a general setting, \( t(x|u)g(u) \). The special stationary linear/Gaussian structure here means this is the same as \( t(u|x)g(x) \). See this by looking at the bivariate distribution for \( (x, u)' \) with vector mean 0 and \( 2 \times 2 \) variance matrix

\[ \Sigma_2 = s \begin{pmatrix} 1 & \phi \\ \phi & 1 \end{pmatrix}, \]

and then noting that both of the conditional distributions - that for \( (x|u) \) and that for \( (u|x) \) - are given by \( t(\cdot|\cdot) \).

Back to the time series context, this means that the process is time reversible, i.e., it looks the same back in time as it does forwards in time. This general result follows directly just by looking at the joint normal distribution, and can be verified by using Bayes theorem: starting with \( (x_t|x_{t-1}) \sim \mathcal{N}(\phi x_{t-1}, v) \) and \( x_{t-1} \sim \mathcal{N}(0,s) \), show using Bayes theorem that \( (x_{t-1}|x_t) \sim \mathcal{N}(\phi x_t, v) \).

### 1.5 Joint Distributions in Compositional/Markov Form

\[ p(x_{1:t}) = p(x_t|x_{t-1})p(x_{t-1}|x_{t-2})\ldots p(x_2|x_1)p(x_1) \]  

(9)

- Uses Markovian (lag 1) structure.
- Each \( p(x_s|x_{s-1}) \) is normal, linear regression from model equation (1).
- Initial value \( x_1 \) (or any other): \( N(0,s) \)

### 1.6 Simulation

What do AR processes look like? Simulate some, for different choices of parameters. How? One way is to just simulate the multivariate normal, but that’s not likely most efficient. Best approach is to utilize as much local structure in the multivariate normal as possible, and here the local structure is Markovian and captured in the compositional decomposition of the density (and corresponding distribution).

Assume we know how to simulate a univariate normal random quantity - we’ll discuss that more later. Simulate a sample path or realization from the AR(1) model of equation (1), generating \( n \) consecutive values \( x_{1:n} \). This is equivalent to both (a) simulating from the full joint normal distribution - the multivariate normal in \( n \)-dimensions of equation (5), and (b) simulating through the sequence of conditional (Markovian) densities in the compositional form equation (9). As follows:
• **Initialize:** draw a sample (a simulated value) \( x_1 \sim p(x_1) \equiv N(0, s) \), via

\[
x_1 = \sqrt{s} z_1
\]

where \( z_1 \sim N(0, 1) \) is a first (synthetic, quasi-random) simulated value in this Monte Carlo exercise.

• For \( t = 2 : n \) in sequence, successively generate \((x_t|x_{t-1})\) from the AR(1) model:

\[
x_t = \phi x_{t-1} + \epsilon_t
\]

where now \( \epsilon_t = \sqrt{v} z_t \) are realized values of the innovations computed from synthetic Monte Carlo draws \( z_t \sim N(0, 1) \), and each evaluated \( x_{t-1} \) is “plugged-in” to the equation for the next time step.
2 Likelihoods and Reference Bayesian Inference in AR(1) Models

Key support material: pages 15-17, 19-23 of the Draft Text Material on time series on the web site are primary.

2.1 Likelihood Functions

Now explicitly recognize dependence on parameters in all densities, so that we have the joint density

\[ p(x_1:n|\theta) = p(x_n|x_{n-1},\theta) p(x_{n-1}|x_{n-2},\theta) \ldots p(x_2|x_1,\theta) p(x_1|\theta). \]  

(10)

In \( \theta \), this gives the likelihood function:

\[ p(x_1:n|\theta) \propto (1 - \phi^2)^{1/2} v^{-n/2} \exp(-Q^*(\phi)/2v) \]  

(11)

with

\[ Q^*(\phi) = Q(\phi) + (1 - \phi^2)x_1^2, \quad Q(\phi) = \sum_{t=2}^{n} (x_t - \phi x_{t-1})^2. \]  

(12)

2.2 Reference Bayesian Analysis Conditioning on Initial Values

For \( n \) large, the effect of the initial value in the complicating factor \( 1 - \phi^2 \) is "small", and the conditional likelihood function is very often used:

\[ p(x_2:n|x_1,\theta) \propto v^{-(n-1)/2} \exp(-Q(\phi)/2v). \]  

(13)

- Approximate analysis relative to actual or “full” likelihood;
- Small difference for large \( n \);
- Valid conditional distribution anyhow;
- Can easily compare with actual likelihood;
- Bayesian analysis via simulation methods will easily allow use of full likelihood, as we shall see.

Key results: Conditional likelihood is of standard form, and MLE and reference Bayesian analysis routine. Reference prior is \( p(\theta) \propto v^{-1} \) (standard reference analysis for normal random samples and linear regressions, of which this is a special case) so that the posterior is easy:

\[ p(\theta|x_1:n) \equiv p(\phi,v|x_1:n) \propto v^{-(n+1)/2} \exp(-Q(\phi)/2v). \]  

(14)

Structure of bivariate posterior for \( (\phi,v|x_1:n) \):

- \( p(\phi|x_1:n,v) \propto \exp(-B(\phi - b)^2/2v) \) with \( B = \sum_{t=1}^{n-1} x_t^2 \) and \( b \equiv \phi_{ML} = B^{-1} \sum_{t=2}^{n} x_t x_{t-1} \). Clearly \( b \) relates to the sample correlation at lag 1, and \( B \) to the sample variance. Then

\[ (\phi|x_1:n,v) \sim N(b,vB^{-1}). \]  

(15)

\( b \) is the reference posterior mode for \( \phi \) as well as the MLE and LSE.

- Verify the quadratic form can be expressed as

\[ Q(\phi) = Q(b) + B(\phi - b)^2. \]  

(16)

This is a special, simple example of the standard decomposition of quadratic forms in linear regression models: \( Q(b) \) is the residual sum of squares based on the “fitted” parameter value \( b \).
• Integrating $\phi$ out of equation (14),

$$p(v|x_{1:n}) \propto v^{-n/2} \exp(-Q(b)/2v).$$

This is the density of an inverse chi-square distribution: $1/v = \kappa/Q(b)$ where $\kappa \sim \chi^2_{n-2}$.

**Gammas and chi-square distributions:** Write $\psi = 1/v$. By transformation $\psi$ has density proportional to $\psi^{(n-2)/2-1} \exp(-\psi Q(b)/2)$, the density of $\psi \sim \text{Ga}((n-2)/2, Q(b)/2)$. By transform again, therefore, $\psi Q(b) \sim \text{Ga}((n-2)/2, 1/2) \equiv \chi^2_{n-2}$ (by definition of the chi-square distribution). Easy to simulate from $p(v|x_{1:n})$: draw $\kappa \sim \chi^2_{n-2}$ and compute $v = Q(b)/\kappa$.

The two distributions $p(\phi|x_{1:n}, v)$ and $p(v|x_{1:n})$ define the joint posterior. Verify that the implied marginal posterior for $(\phi|x_{1:n})$ is a Student T distribution.

**Question and Issue:** If we want to ensure that the model represents a stationary process, then how can we have a normal distribution describing likely values and uncertainty about $\phi$? It may be concentrated in the stationary region, or we may have to condition; but, conditioning destroys the theory just outlined.

### 2.3 Posterior Analysis via Parameter Simulation

If we take the view that much parametric inference in based on posterior sampling (we do), then we care less about specific mathematical forms of marginal posteriors than we do about decomposition of posteriors for Monte Carlo analysis. We can make a draw $(\phi, v)$ from the posterior equation (14) by composition: draw from the margin for $v$ then, conditioning on that value, draw from the conditional for $\phi$ given $v$. Same trick that is used in simulation complicated joint distributions of realizations of the $x_t$ process (which we will do again in the next subsection).

One benefit of using simulation is that we’ll see if sampled $\phi$ values live in the stationary region $|\phi| < 1$. Just by looking at how many fall outside gets us into assessing whether or not the fitted model is consistent with stationarity. If we sample thousands of posterior draws, just looking at the fraction within $|\phi| < 1$ is a start (and often the finish) on this issue. Later we’ll see how this fits into more formal Monte Carlo analysis using accept-reject methods.

- Generate some synthetic AR(1) data and fit the reference Bayesian analysis. Sample the posterior distribution and explore histograms of posterior samples for each of $\phi$ and $v$ (or, better, the innovation scale parameter $\sqrt{v}$).
- Fit the reference analysis to the SOI time series, or any other real data set. Explore posterior histograms, means, etc.

### 2.4 Predictive Simulation: Forecasting and Model Evaluation

Exploring model fit and implications through simulation for any chosen value of $\theta$ is useful, but ignores parameter (estimation) uncertainty. Formally, the relevant distribution for a specified set of unknown future values $x_{(n+1):(n+m)}$ up to $m > 0$ steps ahead is the predictive distribution, with density

$$p(x_{(n+1):(n+m)}|x_{1:n}) = \int p(x_{(n+1):(n+m)}|x_{1:n}, \theta)p(\theta|x_{1:n})d\theta. \quad (18)$$

This can be shown to be a multivariate T distribution, but the use of simulation to generate draws from it is just trivial. Again the key is compositional sampling of a joint distribution:

- Sample a synthetic parameter value from $p(\theta|x_{1:n})$,
- plug-in this generated value for $\theta$ in its placeholder in the conditioning of $p(x_{(n+1):(n+m)}|x_{1:n}, \theta)$, and then
• sample $x_{(n+1):(n+m)} \sim p(x_{(n+1):(n+m)} | x_{1:n}, \theta)$.

We have already seen this last step - sampling from the model $AR(1|\theta)$ for a given parameter value, again by composition now sequencing through Monte Carlo draws of $(x_{n+1}|x_n, \theta)$, then $(x_{n+2}|x_{n+1}, \theta)$, and so on. Note that now, however, we must conditional on the pre-initial value $x_n$ to get started, as that as been observed.

Sample realizations - “simulated futures” of the process - generated this way are draws from the formally correct posterior predictive distribution for the evolution of the process into the future. They are based on and incorporate the information relevant to estimation of the parameters, but now also reflect estimation uncertainty: if there is a lot of uncertainty about $\theta$, due to small $n$ for example, then the resulting higher level of uncertainty in the posterior feeds through to the predictions.

• Explore “simulated futures” based on models fitted to synthetic AR(1) data and real data, such as the SOI series.

3 Two Classes of Hidden Markov Models with AR(1) Components

3.1 AR(1) Process Observed with Noise

Observed values of a process are now $y_t$, and

\[
\begin{align*}
  y_t & = x_t + \nu_t \\
  x_t & \leftarrow AR(1|\theta)
\end{align*}
\]

where $\nu_t \sim N(0, w)$ and with $\nu_t \perp \nu_s$ and $\nu_t \perp \epsilon_s$ for all $t, s$. The $\nu_t$ terms are errors of measurement, or of observation, that “corrupt” the signal $x_t$.

This is a hidden Markov model (HMM), and one of the simplest. It is also one of the more important models in time series as many real processes are not directly observable: measurement error and other forms of technical error, noise obscure the signal $(x_t)$. The first equation is equivalent to $(y_t|x_t) \sim N(x_t, w)$, so that $y_t$ is an unbiased measurement, but not perfect. The relative values of $w$ and $v$ define signal-to-noise characteristics. In a stationary model we know $V(x_t) = s = v/(1 - \phi^2)$ so that the variance of each observation is $s + w$ and the SNR (signal-to-noise ratio) is $s/(s + w)$.

What are the stochastic characteristics of the observed process $y_t$? The $y_t$ process is stationary and linear, Gaussian so each $y_t$ has a normal marginal distribution, each pair is bivariate normal, and so forth. Is it Markovian? Is it time reversible?

3.2 A Class of Stochastic Volatility Models

A famous class of HMMs are stochastic volatility (SV) models (SVMs) of interest in quantitative finance, the cornerstone of many statistical decision support tools in financial research and investment management. The canonical example:

\[
\begin{align*}
  y_t & \sim N(0, \sigma_t^2) \\
  \sigma_t & = \exp(\mu + x_t) \\
  x_t & \leftarrow AR(1|\theta)
\end{align*}
\]
• Fix parameters $\theta = (\phi, v)$ and $\mu$ and initial state $x_0$
• Simulate sample paths (“trajectories”) of $x_t, y_t$
• Structure in $x_t$ series
• Structure in $y_t$ series
• Structure in $y_t^2$ series
• Real data: Returns on international exchange rate markets, where all the action is in the changes in variance of returns, and it is important to model and capture persistence in variances (volatilities).
• Add a dynamic mean: Econometrics and dynamic models - small drifts
• Fit models to data: Assessment, prediction

Some insight into the structure of the model, and also complications in model fitting, arise by noting the implication that, conditional on $\mu, x_t$, the random quantity $y_t^2 = \sigma_t^2 \kappa_t$ where $\kappa_t \sim \chi^2_1$. So if $z_t = \log(y_t^2)/2 = \log(|y_t|)$,

\[
\begin{align*}
z_t & = \mu + x_t + \nu_t \\
x_t & \leftarrow AR(1|\theta)
\end{align*}
\]

where $\nu_t = \log(\kappa_t)/2$. So we have a model in which the observed quantities, now $z_t$, are a constant (intercept, level, defining the baseline volatility on the log scale) plus a latent AR(1) process $x_t$ that defines the time-correlated changes in volatility. This is very similar to the linear-Gaussian AR(1) plus noise model, the simple HMM, but now with two differences: (a) the model has an intercept term, which is only a small detail; (b) the observational noise is non-Gaussian, a bigger issue - the distribution of $\nu_t$ is that of 1/2 times the log of a chi-square on 1 degree of freedom, which is not quite Gaussian, somewhat skewed, and a little difficult to work with mathematically. One approach is to approximate this with a normal distribution, and that puts us into the linear, Gaussian hidden AR model framework. More effective approximations exist and can be developed using simulation methods; this model context is one in which we really begin to need Monte Carlo methods for model analysis and prediction.

Multivariate normal distribution theory underlies a good part of the structure and analysis of these models - hidden Markov models with latent AR components, and also the SV models even though the sampling distribution is non-normal. The full joint distribution of any set of log volatilities is, initially, multivariate normal as it is just a linear AR(1) process, and much of the trickery needed in model fitting, especially using MCMC methods, relies very heavily on the implied collections of marginal and conditional distributions. So we do need to know normal distribution theory intimately.