1 Autoregressive Models

1.1 Introduction

AR\((p)\) models for univariate time series are Markov processes with dependence of higher order than lag-1 in the univariate state space. Linear, Gaussian models represent a practically important class of models for time series analysis and decomposition of processes into components, and also an important class of models within which to explore and understand the ideas of higher-order dependence in Markov processes. Refer to *** Support notes on Linear Processes, AR models etc *** on course web site - mainly the first several pages of chapter 2 of that note – , and Chapters 9 and 15 of West & Harrison Bayesian Forecasting and Dynamic Models.

1.2 AR\((p)\) Model Form

Data \(y_t\) are generated from an AR\((p)\) model

\[
y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + \epsilon_t \quad \text{where} \quad \epsilon_t \sim N(0, v)
\]

for \(t = 1, 2, \ldots\) (and conceptually \(t = 0, -1, \ldots\)) and with \(\epsilon_t \perp \epsilon_s\) for \(t \neq s\).

- **Notation**: \(y_t\) is used in place of earlier \(x_t\) for notation, and the reason will become clear.
- **Notation**: \(y_t \leftarrow AR(p|\theta)\) with \(\theta = (\phi, v)\) and \(\phi = (\phi_1, \ldots, \phi_p)'\).
- \(y_t\) is regressed on \(p\) past, most recent, consecutive values of the process.
- Linear prediction model.
- Obvious extension of AR\((1)\) models.
- Homogenous Markovian model - the same model applies for all \(t\), since the parameters \((\phi, v)\) are constant in time.
- \(y_t\) is a linear, homogenous Gaussian process, and is time reversible. Other models with non-Gaussian innovations are linear but not reversible.

1.3 Backshift Operators and Characteristic Polynomials

- Operator \(B\) such that \(By_s = y_{s-1}\) and so \(B^k y_s = y_{s+k}\) for all \(k\).
- \(y_t = \sum_{j=1}^{p} \phi_j B^j y_t + \epsilon_t\)
- \(\Phi(B)y_t = \epsilon_t\)
- **Characteristic Polynomial** \(\Phi(u) = 1 - \phi_1 u - \phi_2 u^2 - \cdots - \phi_p u^p\). This is a polynomial of order \(p\) defined on \(|u| < 1\).

1.4 Inversion and Moving Averages (MAs)

As in the AR\((1)\) case, iterative substitution of \(y_{t-1}\) then \(y_{t-2}\) and so on in the right hand side of the model equation represents \(y_t\) as a linear function of more distant past \(y_s\) and \(\epsilon_t, \epsilon_{t-1}, \ldots\). Formally, assuming the inversion can be performed,

\[
y_t = \Phi(B)^{-1}\epsilon_t = \Pi(B)\epsilon_t
\]

or

\[
y_t = \epsilon_t + \pi_1 \epsilon_{t-1} + \pi_2 \epsilon_{t-2} + \cdots
\]
where the (likely infinite order) polynomial \( \Pi(u) = 1 + \pi_1 u + \pi_2 u^2 + \cdots \) satisfies \( \Phi(u) \Pi(u) = 1 \).

This gives the explicit representation of \( y_t \) as a linear combination of independent innovations, a linear (Gaussian) process. This is also referred to as a Moving Average (MA) representation. If the weights \( \pi_j \) cut-off to zero after some finite lag \( q \), we have a finite MA\((q)\) representation.

1.5 Stationarity

A stationary AR\((p)\) process is such that this inversion exists. The moving average weights must decay to zero eventually, otherwise the linear combination of past innovations will explode. If the representation exists, then evidently

- \( E(y_t) = 0 \) for all \( t \),
- \( V(y_t) = v \sum_{j=0}^{\infty} \pi_j^2 \) for all \( t \), and this shows that the weights must decay rapidly with \( j \),
- \( \text{Cov}(y_t, y_{t-k}) = \gamma(k) \) is some function of the \( \pi_j \) weights, but depends only on lag \( k \) and not on \( t \).
- These moments and all other properties are determined by the values of \( p, \phi, v \).

1.6 Characteristic Roots

Write \( \Phi(u) = \prod_{j=1}^{p} (1 - \alpha_j u) \) where the \( \alpha_i \) are the reciprocals of the roots of the characteristic polynomial, i.e., \( \Phi(\alpha_j^{-1}) = 0 \) for each \( j = 1, \ldots, p \).

- \( \Pi(u) = \Phi(u)^{-1} = \prod_{j=1}^{p} (1 - \alpha_j u)^{-1} \) exists on \( |u| < 1 \) if, any only if, \( |\alpha_j| < 1 \) for each \( j = 1, \ldots, p \).
- \( \{\alpha_j\} \) are the characteristic roots of the AR\((p)\) model.
- Polynomial roots may be real or complex. Complex roots occur in pairs of complex conjugates.
  - A real root is any number in \( -1 < r < 1 \).
  - A pair of complex conjugate roots has the form \( \alpha, \alpha^* = r \exp(\pm i\omega) \) for some modulus \( r \), such that \( 0 < r < 1 \), and argument (or angle) \( \omega \) such that \( -\pi < \omega < \pi \) (or, in some conventions, \( 0 < \omega < 2\pi \)).
  - Matlab: \texttt{abs} and \texttt{angle}. (R/Plus: \texttt{Mod} and \texttt{Arg}).
  - \( r \exp(\pm i\omega) = r \cos(\omega) \pm ir \sin(\omega) \).
  - \( \alpha + \alpha^* = 2r \cos(\omega) \) and \( \alpha\alpha^* = r^2 \).
  - The period, or wavelength, of the complex roots is \( 2\pi/\omega \).

1.7 AR(2) Examples

When \( p = 2 \), \( \Phi(u) = 1 - \phi_1 u - \phi_2 u^2 = (1 - \alpha_1 u)(1 - \alpha_2 u) \) may have either two, distinct real roots or one pair of complex conjugate roots. Note that \( \phi_1 = \alpha_1 + \alpha_2 \) and \( \phi_2 = -\alpha_1\alpha_2 \). If the roots are real, the autocorrelations in the process are a composition of those of two AR\((1)\) processes with parameters given by the two real roots, and in fact the AR\((2)\) model can be represented as the sum of two such AR processes, as we shall see below in discussion of model decompositions.

If the roots are complex, the AR\((2)\) process is quasi-periodic. Sample trajectories have the appearance of “noisy” damped cosine waves of fixed wavelength \( 2\pi/\omega \). The “noise” is random variation in amplitude and phase of the waveform through time, injected by the innovations sequence. The damping is induced by the modulus \( r \). A low modulus rapidly damps the waveform between time points, prior to the injection of the next innovation. A very persistent waveform, closer to a sinusoidal form, is generated in cases of \( r \) close to unity.
Model decomposition theory below shows how all AR\((p)\) models can be decomposed, and hence understood both theoretically and from a quantitative, practical viewpoint, in terms of basic AR(1) and (almost) AR(2) processes.

### 1.8 Partial Inversion of AR Models

Sometimes we use fit higher-order models and then may interpret part of the structure as arising from a lower-order model with correlated innovations, i.e., ARMA processes.

For example, suppose \(p = 3\) and we have one real root \(\rho\) and a pair of complex conjugate roots \(\alpha, \alpha^*\) so that \(\Phi(u) = (1 - \rho u)(1 - \alpha u)(1 - \alpha^* u)\). Suppose that \(\rho\) is fairly small so that \((1 - \rho u)^{-1} \approx 1 + \rho u + \rho^2 u^2\) for \(|u| < 1\). Then can rewrite the AR model \(\phi(B)y_t = \epsilon_t\) as

\[
(1 - \alpha B)(1 - \alpha^* B)y_t \approx (1 + \rho B + \rho^2 B^2)\epsilon_t
\]

or

\[
y_t \approx \phi'_1 y_{t-1} + \phi'_2 y_{t-2} + \epsilon_t + \pi_1 \epsilon_{t-1} + \pi_2 \epsilon_{t-2}
\]

where \(\phi'_1 = 2r \cos(\omega), \phi'_2 = -r^2, \pi_1 = \rho\) and \(\pi_2 = \rho^2\).

- \(y_t\) looks like an ARMA(2,2) process - an AR(2) process in which the innovations are correlated and themselves have a second-order structure.
- Obvious extensions to ARMA\((p, q)\) models are of interest.
- In this example with \(\rho\) quite small, the second term may be negligible and the process looks approximately like an ARMA(2,1) process.
- Very often, higher-order AR\((p)\) models have underlying, lower-order AR structure that is somewhat obscured by measurement or timing errors that induce correlation between the innovations, and can be recovered through the device of fitting a higher-order model and then using this idea of partial inversion, at least at an exploratory level.

### 1.9 Reference Conditional Linear Model for Inference and Prediction

Refer back to Homework 3, Question 5, where you fitted AR\((p)\) models to the SOI series and explored inference in the reference analysis.

Under this model, conditioning on the set of initial values \(y_1, \ldots, y_p\) and assuming we observe \(n > 2p\) consecutive values, we have a model that has the form of a linear regression, and reference Bayesian inference is standard. With \(y_{(p+1):n} = (y_{p+1}, y_{p+2}, \ldots, y_n)'\) and \((n - p) \times p\) autoregressive design matrix

\[
H = \begin{pmatrix}
y_p & y_{p-1} & \cdots & y_1 \\
y_{p+1} & y_p & \cdots & y_2 \\
\vdots & \vdots & \ddots & \vdots \\
y_{n-1} & y_{n-2} & \cdots & y_{n-p}
\end{pmatrix},
\]

we have the linear model expression

\[
y_{(p+1):n} = H\phi + \epsilon_{(p+1):n} \quad \text{with} \quad \epsilon_{p+1:n} \sim N(0, vI).
\]

The conditional reference posterior for \(\theta\) has the compositional form
\[
(\phi|v, y_{1:n}) \sim N(b, vB^{-1}), \\
(v^{-1}|y_{1:n}) \sim Ga((n - 2p)/2, Q(b)/2),
\]
with
\[
Q(\phi) = (y_{(p+1):n} - H\phi)'(y_{(p+1):n} - H\phi), \quad B = H'H \quad \text{and} \quad b = B^{-1}H'y_{(p+1):n}.
\]

Here \( b \) is the conditional MLE, LSE and reference posterior mean and mode for \( \phi \), and \( Q(b) \) is the residual sum of squares from the conditional regression analysis.

This is all (trivially) coded in the existing little Matlab functions. They also provide for posterior simulation - direct sampling from this conditional reference posterior, and simulation of predictive distributions - “sampling the future”. Again, refer back to Homework 3, Question 5 and the analysis support code used there.

1.10 Linear Systems: State Space Representation

Introduce the \( p \)-dimensional state vector \( x_t = (y_t, y_{t-1}, \ldots, y_{t-p+1})' \) for all \( t \). Then the AR(\( p \)) model may be re-expressed as

\[
y_t = F'x_t \\
x_t = Gx_{t-1} + F\epsilon_t
\]

where

\[
F = \begin{pmatrix}
1 \\
0 \\
0 \\
\vdots \\
0
\end{pmatrix} \quad \text{and} \quad G = \begin{pmatrix}
\phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\
1 & 0 & \cdots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cdots & 1 & 0
\end{pmatrix}.
\]

Here \( G \) is the state evolution, or transition matrix in the extended state space representation. Note that this maps the state from one to \( p \) dimensions and so converts the \( p^{\text{th}} \) order Markovian dependence to a first order dependence.

- This representation is one reason for the notational use of \( y_t \) for data, since now \( x_t \) is the \( p \)-vector state variable.
- An easy extension to a latent AR process - HMM with the underlying hidden state being a higher-order model - is given by the “AR(\( p \)) in noise” extension in which \( y_t = F'x_t + \nu_t \).

1.11 Forecast Function and Eigenstructure

Insight into the dependence structure is generated by inspection of the forecast function \( f_t(k) = E(y_{t+k}|y_{1:t}) \) of the process as a function of the “look-ahead” horizon \( k = 1, 2, \ldots \). It is easily seen that \( E(x_{t+k}|y_{1:t}) = G^kx_t \) so that \( f_t(k) = F'G^kx_t \).

Now \( G \) is a square matrix. Assume that \( G \) has distinct, non-zero eigenvalues - this will be almost surely the case when a \( \phi \) is derived from a model fit to real data. The \( G \) has an eigendecomposition \( G = E\Lambda E^{-1} \) where the \( p \times p \) eigenvector matrix \( E \) has columns that are the eigenvectors of the corresponding elements of the diagonal matrix \( \Lambda \). That is, \( G\lambda_j = e_j\lambda_j \) where \( E = [e_1, \ldots, e_p] \) and \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p) \). The
eigenvalues and vectors can be real or complex valued. Since $G$ is real valued, any complex eigenvalues must occur in conjugate pairs. Then $G^k = E \Lambda^k E^{-1}$ and so

$$f_t(k) = \sum_{j=1}^{p} c_{t,j} \lambda_j^k$$

for some (real or complex) numbers $c_{t,j}$ that depend on $E$ and $x_t$.

- A real valued eigenvalue $\lambda_j$ contributes a term $\lambda_j^k$ to the forecast function. This will explode unless $|\lambda_j| < 1$, and will otherwise decay to zero with increasing $k$.

- A pair of complex conjugate eigenvalues $r \exp(\pm i \omega)$ will contribute terms proportional to

$$r^k \exp(\pm i \omega k) = r^k \{\cos(\omega k) \pm i \sin(\omega k)\}.$$

Since the forecast function is real the two terms must have canceling imaginary components, so leading to a term proportional to

$$r^k \cos(\omega k + a_t)$$

for some phase $a_t$ that depends on the current state $x_t$. Unless $|r| < 1$ this will also be explosive; if $|r| < 1$, then the form is a damped cosine wave of fixed wavelength $2\pi/\omega$ and current state dependent phase.

- $y_t$ is a linear combination of these processes.

- With models of higher order $p$, many such component processes exist. Some may be of real practical (physical, economic) significance and interest, whereas some - often those of very low moduli, will often represent high frequency or short-term noise.

- Inference on $(\phi, v)$ (and $p$) leads to inference on the eigenstructure of the model and hence on the underlying components in the time series. Simulation of the posterior for $(\phi, v)$ leads easily to simulated values of the eigenvalues and so forth. Plug-in estimates of $\phi$, such as the reference posterior mean $b$, provide a start.

**Key result:** It can be shown (with some interesting linear algebra) that the eigenvalues $\lambda_j$ are precisely the characteristic roots: $\lambda_j = \alpha_j$ for $j = 1, \ldots, p$. Hence a stationary AR($p$) is characterized by a $G$ matrix that has all eigenvalues of less than unit modulus, whether real or complex. In such cases, the forecast function is clearly “well-behaved”, with the components damping out as $k$ increases, eventually leading to $f_t(k) \rightarrow 0 = E(y_{t+k})$ as $k$ increases.

### 1.12 Autocorrelations

In the state space representation, $E(x_t) = 0$ and $V(x_t) = S$ where $S$ satisfies $S = GSG' + U$ and $U$ is the $p \times p$ matrix of zeros except for $U_{1,1} = v$. We can see the form of the a.c.f. of $y_t$ easily. Since $y_t = x'_t F$ we have

$$\gamma(k) = E(y_{t+k}y_t) = F' E(x_{t+k}x_t') F.$$

Now $x_{t+k} = G^k x_t +$ terms involving $\epsilon_{t+1}, \ldots, \epsilon_{t+k}$. Hence $E(x_{t+k}x_t') = G^k S$ and so

$$\gamma(k) = F' G^k S F = \sum_{j=1}^{p} g_j \lambda_j^k$$

for some constant $g_j$ that depend on $(\phi, v)$. As a result, the autocorrelations $\rho(k)$ have the same form as a function of lag $k$; that is, precisely the same form as the forecast function. Autocorrelations of AR($p$) processes are a mixture of damped AR(1)-like terms that decay exponentially (real eigenvalues) and may oscillate (real negative eigenvalues), and damped AR(2)-like cosine forms (complex conjugate pairs of eigenvalues).
1.13 AR Model and Process Decompositions

The eigentheory is completed with an explicit representation of $y_t$ in terms of underlying (latent, but identifiable) components of AR(1) and/or AR(2) (actually ARMA(2,1)) forms.

- Define a transformed $p-$vector state variable $z_t = E^{-1} x_t$. Also set $F_o = E' F$ and $F_e = E^{-1} F$.
- Then $y_t = F'_o z_t$ and $z_t = \Lambda z_{t-1} + F_e \epsilon_t$ for all $t$.
- Elements of $z_t = (z_{t,1}, z_{t,2}, \ldots, z_{t,p})'$ are individual AR(1) processes, with $z_{t,j}$ having AR coefficient $\lambda_j$. They are correlated since they are driven by the same innovations.
- Real eigenvalues lead to real components: $z_{t,j} \leftarrow AR(1|\lambda_j, v_j)$.
- Complex eigenvalues $r \exp(\pm i\omega)$ lead to pairs of complex and conjugate AR(1) processes. The linear combination of two such processes must be real, and this leads to a real component that has the structure of an ARMA(2,1) component in which the AR(2) part is damped by $r > 0$, and quasi-periodic with fixed frequency $\omega$, i.e., wavelength $2\pi/\omega$, but time-dependent amplitude and phase.

See the Time Series support notes, especially the discussion of AR(2) decompositions, and the also Chapters 9 - especially - and 15 of West & Harrison Bayesian Forecasting and Dynamic Models for further details.
2 Sequential Learning and TV AR models

2.1 Sequential Representations of Learning in Autoregressive Models

Investigation of the details of how posterior distributions for \((\phi, v)\) are sequentially updated as new data arises is useful for, among other things, defining a framework for extension to time-varying parameter models.

We can recast the model in the following regression form. With the \(p\)-dimensional state vector \(x_t = (y_t, y_{t-1}, \ldots, y_{t-p+1})'\) for all \(t\) we have

\[ y_t = x_{t-1}' \phi + \epsilon_t \]

and, under the regression analysis as already described, the posterior for \(\theta = (\phi, v)\) has the conjugate normal-inverse gamma form (see §1.9). We need a change of notation, and now write the posterior as

\[
\begin{align*}
(\phi|v, y_{1:t}) &\sim N(m_t, vM_t), \\
(v^{-1}|y_{1:t}) &\sim Ga(n_t/2, n_t s_t/2).
\end{align*}
\]

This conditional posterior is valid for all times \(t\), and so the defining quantities \(\{m_t, M_t, n_t, s_t\}\) are naturally related as time \(t\) varies. In particular, the “time \(t\) update” involves the mapping from their values at \(t-1\) - based on data \(y_{1:(t-1)}\) - to their values at \(t\), representing the additional information provided by the time \(t\) observation \(y_t\).

The following key theory defines the sequential updating, and is quite general.

Suppose \(p(\phi, v|y_{1:(t-1)})\) has the above normal-inverse gamma form with defining parameters

\[
\{m_{t-1}, M_{t-1}, n_{t-1}, s_{t-1}\}
\]

Then:

- The one-step ahead forecast distribution conditional on \(v\) is

\[ (y_t|y_{1:(t-1)}, v) \sim N(x_{t-1}' m_{t-1}, q_t v) \]

with \(q_t = 1 + x_{t-1}' M_{t-1} x_{t-1} \).

- The time \(t\) posterior distribution is \(p(\phi, v|y_{1:t}) = p(\phi|v, y_{1:t})p(v|y_{1:t})\) and is normal-inverse gamma with parameters \(\{m_t, M_t, n_t, s_t\}\) that are computed as follows:

\[
\begin{align*}
- m_t &= m_{t-1} + A_t e_t, \\
- M_t &= M_{t-1} - A_t A_t' q_t, \\
- n_t &= n_{t-1} + 1, \\
- s_t &= (n_{t-1} s_{t-1} + e_t^2/q_t)/(n_{t-1} + 1).
\end{align*}
\]

with

\[
\begin{align*}
- e_t &= y_t - x_{t-1}' m_{t-1}, \text{ the one-step ahead forecast error, and} \\
- A_t &= M_{t-1} x_{t-1}/q_t, \text{ the adaptive coefficient } p-\text{vector,}
\end{align*}
\]

These results flow from standard normal theory and Bayes’ theorem (see Multivariate Normal Theory notes).

Some comments and alternative expressions are of interest:

- The update for \(m_t\) is a “predictor/corrector” form: The prior or “predicted” value for \(\phi\), namely \(m_{t-1}\), is corrected by the weighted forecast error. A large forecast error implies a large correction, and vice-versa.
• $M_t < M_{t-1}$, so that we apparently always gain information.
• The degrees of freedom $n_t$ increases by one per observation.
• The error variance estimate $s_t$ is updated as a weighted average of the predicted estimate $s_{t-1}$ and
  the forecast error scaled by $q_t$. A larger forecast error leads to an inflation of the estimate of error variance.
• All conditional normal distributions convert to the corresponding $T$ distributions on marginalization
  over $v$. For example,
  
  - $(y_t|y_1:(t-1))$ is univariate $T$ on $n_{t-1}$ degrees of freedom, with mean $x'_{t-1}m_{t-1}$ and scale $q_ts_{t-1}$,
    whereas
  - the posterior $p(\phi|y_1:t)$ is $p-$variate $T$ distributed on $n_t$ degrees of freedom with mean $m_t$ and
    scale matrix $M_ts_t$.
• Very important alternative representations of $\{m_t, M_t\}$ are the forms derived directly from Bayes’
  theorem, namely

\[
 m_t = M_t(M_{t-1}^{-1}m_{t-1} + x_{t-1}y_t) \quad \text{and} \quad M_t^{-1} = M_{t-1}^{-1} + x_{t-1}x'_{t-1}.
\]

These are standard formulæ, though the new, alternative representations above are both computationally more
  efficient and numerically more stable as no matrix inversions are required. Some additional practical points
  related to initialization:
• Since the updating results hold true for all $t$, it is clear that we can now consider analysis based on any
  initial prior of the normal-inverse gamma form at $t = p$. We need to consider the initial point as $t = p$
  since the required regression vector $x_t$ is available only for $t \geq p$. That is, analysis can be initialized
  at any specified values of $\{m_p, M_p, n_p, s_p\}$ prior to implementing the sequential form of the analysis
  beginning with the “first” observation at $t = p + 1$.
• An alternative initialization involves fitting the reference posterior distribution based on an initial set
  of $q > p$ observations, and then beginning the sequential analysis at a “first” time point $t = q + 1$ with
  $\{m_q, M_q, n_q, s_q\}$ defined by the reference posterior.

### 2.1.1 Useful Linear Algebraic Identities

For reference, the latter equation and the alternative form of $M_t$ are related to a key and broadly useful matrix
  identity, in this case - dropping the time indices for generality and clarity - simply

\[
 (M^{-1} + xx')^{-1} = M - Mxx'M/(1 + x'Mx)
\]

for any positive definite and symmetric $p \times p$ matrix $M$ and $p-$vector $x$.

A more general version involves a $p \times q$ matrix $X$ and a $q \times q$ positive definite symmetric matrix $V$, when the identity is

\[
 (M^{-1} + XV^{-1}X')^{-1} = M - MX(V + X'MX)^{-1}X'M.
\]

### 2.2 Model Likelihood Function

One most useful side product of the sequential analysis is an easy evaluation of the model likelihood - the
  joint density of observations unconditional on model parameters. Simply note that, at each time $t$ we obtain
the one-step ahead forecast, or predictive density \( p(y_t|y_{1:(t-1)}) \). This is the univariate T p.d.f.

\[
p(y_t|y_{1:(t-1)}) = c(n_{t-1})(s_{t-1}q_t)^{-1/2} \left( 1 + \frac{e_t^2}{n_{t-1}s_{t-1}q_t} \right)^{-(n_{t-1}+1)/2}
\]

where \( c(\nu) \) is the constant of normalization of the T density on \( \nu \) degrees of freedom,

\[
c(\nu) = \frac{\Gamma((\nu + 1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}}.
\]

Then, the joint p.d.f. of any set of observations \( y_{k:n} \) is given by composition as

\[
p(y_{k:n}|y_{1:k-1}) = \prod_{t=k}^n p(y_t|y_{1:(t-1)}).
\]

For example, if we start with \( t = p + 1 \) and some initial values of parameters \( \{m_p, M_p, n_p, s_p\} \) as discussed above, then \( k = p + 1 \) - the above expression defines the joint density of the data from thereon.

The value of the joint density function is also named the marginal likelihood of the model. For example, we may rerun the analysis at different values of the model order \( p \), and then \( p(y_{1:n}) \) is actually \( p(y_{1:n}|p) \), the value of the data density conditional on that value of \( p \). As we change \( p \) and rerun the analysis, this maps out the likelihood function over \( p \) for assessment of model order. A prior distribution over \( p \) might then be used to convert to a posterior distribution for model order, for example.
2.3 TVAR Models

Some non-stationary phenomena can be regarded as “locally stationary” in the sense that a standard model, such as an AR\((p)\) model, provides an adequate and useful functional form for the data generating process, but its adequacy relies on permitting the defining parameters to take different values over time. This simple concept is in fact quite powerful in some applications, and in fact underlies very widely used methods of local smoothing in time series in many areas of social and natural sciences, engineering signal processing, and short-term forecasting in business and economics. The comprehensive framework of dynamic state space models (West and Harrison, 1997) demonstrates the broad scope of applicability of the concept, as well as various models classes.

A core class of such non-stationary models is the class of Time-Varying Autoregressive Models (TVAR), a rather simple but nevertheless very useful extension of AR models.

2.3.1 Concept: Random Drift in Parameters

The key concept is simply to allow model parameters to vary randomly in time, according to a stochastic process model. We will examine just random walk models in detail, but the concept is more general. The AR\((p)\) model \(y_t \leftarrow AR(p)(\phi, v)\) is extended to the TV AR\((p)\) model in which, at any time \(t\),

\[
y_t = x_{t-1}' \phi_t + \epsilon_t \equiv \sum_{j=1}^{p} \phi_{t,j} y_{t-j} + \epsilon_t
\]

where \(\phi_t = (\phi_{t,1}, \phi_{t,2}, \ldots, \phi_{t,p})'\) is the - now - time-varying AR parameter vector. Any model for time variation might be considered. A simple random walk model has several points of recommendation. A such model has the form

\[
\phi_t = \phi_{t-1} + \omega_t
\]

where \(\omega_t\) is a sequence of zero-mean \(p\)-vector variates with \(\omega_t \perp \perp \omega_s\) for \(t \neq s\). The \(\omega_t\) variates constitute a sequence of random “shocks” to the model that define the temporal evolution of the AR parameters.

This random walk parameter evolution is simple and has the attraction that it is neutral with respect to directions of change in parameters, and allows them to “wander” freely over time so permitting substantial change in model form over long periods.

2.3.2 Sequential Learning in TVAR Models

Suppose that, at time \(t-1\), information on the current AR parameter is summarized via \((\phi_{t-1}| v, y_1;(t-1)) \sim N(m_{t-1}, vM_{t-1})\) while \((v^{-1}|y_1;(t-1)) \sim Ga(n_{t-1}/2, n_{t-1}s_{t-1}/2)\) just as in the static parameter model. Then:

- Assume that \(\phi_t = \phi_{t-1} + \omega_t\) where

\[
(\omega_t| v, y_1;(t-1)) \sim N(0, vW_t)
\]

independently of \(\phi_{t-k}\) for \(k \geq 1\). Note that we include \(v\) as a constant factor in the variance of \(\omega_t\) for consistency throughout.

- It easily follows that the “evolution” of the parameter to time \(t\) results in a (prior) distribution for the changed value that is just

\[
(\phi_t| v, y_1;(t-1)) \sim N(m_{t-1}, vM_{t|t-1})
\]

where \(M_{t|t-1} = M_{t-1} + W_t\).
Observing $y_t$, $p(\phi_t|y_t, y_{1:t}) \sim N(m_t, v M_t)$ where the update equations are precisely as in the AR model but now the initial distribution has a variance matrix component $M_{t|t-1}$ in place of $M_{t-1}$.

The update for $p(v|y_{1:t})$ is also essentially as in the AR model but again with the small change that $M_{t|t-1}$ replaces $M_{t-1}$.

In summary, the sequential learning proceeds as follows:

1. The time $t - 1$ posterior is parameterized by $\{m_{t-1}, M_{t-1}, n_{t-1}, s_{t-1}\}$;
2. The one-step ahead prior at $t - 1$ is
   \[
   (\phi_t|v, y_{1:(t-1)}) \sim N(m_{t-1}, v M_{t|t-1}),
   \]
   \[
   (v^{-1}|y_{1:(t-1)}) \sim \text{Ga}(n_{t-1}, n_{t-1}s_{t-1});
   \]
3. The one-step ahead forecast distribution conditional on $v$ is
   \[
   (y_t|y_{1:(t-1)}, v) \sim N(x_{t-1}'m_{t-1}, q_tv)
   \]
   with $q_t = 1 + x_{t-1}'M_{t|t-1}x_{t-1}$.
4. The time $t$ posterior distribution is $p(\phi, v|y_{1:t}) = p(\phi|v, y_{1:t})p(v|y_{1:t})$ and is normal-inverse gamma with parameters $\{m_t, M_t, n_t, s_t\}$ that are computed as follows:
   - $m_t = m_{t-1} + A_t e_t$,
   - $M_t = M_{t|t-1} - A_t A'_t q_t$,
   - $n_t = n_{t-1} + 1$,
   - $s_t = (n_{t-1} s_{t-1} + e_t^2/q_t)/(n_{t-1} + 1)$.

with
   - $e_t = y_t - x_{t-1}'m_{t-1}$, the one-step ahead forecast error, and
   - $A_t = M_{t|t-1}x_{t-1}/q_t$, the adaptive coefficient $p-$vector.

Again, these results flow from standard normal theory and Bayes’ theorem (see Multivariate Normal Theory notes). Also, again exactly as in the static AR case, important alternative representations of $\{m_t, M_t\}$ are the forms derived directly from Bayes’ theorem, namely

\[
m_t = M_t(M_{t|t-1}^{-1}m_{t-1} + x_{t-1}'y_t) \quad \text{and} \quad M_t^{-1} = M_{t|t-1}^{-1} + x_{t-1}x_{t-1}'.
\]

Note that the AR($p$) model is, of course, recovered as the special case in which $W_t = 0$ so that $M_{t|t-1} = M_{t-1}$. Otherwise, the model now allows for parameter change through non-zero $W_t$ matrices. Controlling the degree of change is very often desirable, in order that a model parametrization change smoothly consistent with scientific context. Within a Gaussian framework, this involves care in specifying or controlling the estimation of these variance matrices.
2.3.3 Concept: Random Parameter Change as Loss of Information

Between times points \( t - 1 \) and \( t \), the increased in variance from \( M_{t-1} \) to \( M_{t|t-1} = M_{t-1} + W_t \) reflects a decreased precision, i.e., loss of information. This concept of information loss is key to a parsimonious and efficient method of structuring parameter change models - the notion of information discounting. One specific application of the more general concept of variance matrix discounting (West and Harrison, 1997, chapter 6) to structure and specify such random-change models - in fact, the simplest such approach - is to assume a constant rate of loss of information about parameters over time. For a fixed, specified discount factor \( \delta \) with \( 0 < \delta < 1 \), specify

\[
W_t = M_{t-1}(\delta^{-1} - 1).
\]

That is, the loss of information \( W_t \) is a (usually rather small) fraction of the existing information \( M_{t-1} \). For instance, \( \delta = 0.99 \) implies a per period loss of information of about 1%, whereas \( \delta = 0.9 \) leads to information attrition at a rate of about 11% per period.

One key implication is that, simply,

\[
M_{t|t-1} = M_{t-1}/\delta,
\]

the discount factor inducing a (usually small) inflation in the elements of the variance matrix between time points. The discount factor \( \delta \) can be specified, and analysis repeated with differing values. The theory of sequential/compositional computation of the model likelihood applies directly to provide for assessment of different values along with the model order \( p \).