Time Series Analysis and Forecasting

Stephen Vardeman
Analytics Iowa LLC

June 3, 2013

Abstract

These notes summarize the main points of an MS-level statistics course in time series analysis and forecasting. Material here has been drawn from a variety of sources including especially the books by Brockwell and Davis, the book by Madsen, and the book by Cryer and Chan.

Contents

1 Notation, Preliminaries, etc.

The course is about the analysis of data collected over time. By far the best-developed methods for such data are appropriate for univariate continuous observations collected at equally spaced time points, so that simply indexing the observations with integers and talking about "time period t" is sensible. This is where we'll begin. So we'll consider a (time) series of values

$$y_1, y_2, \ldots, y_n$$

and write

$$\underset{n\times 1}{\boldsymbol{Y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{1}$$

Sometimes the purpose of time series analysis is more or less "scientific" and amounts to simply understanding interpretable structure in the data. But by far the most common use of time series methods is *predicting/forecasting*, and very/most often the motivating application is economic in nature. We'll feature forecasting in this edition of Stat 551 and pay special attention to methods and issues that arise in the practice of such forecasting. We begin with some basic notation and ideas.

1.1 Linear Operations on Time Series

Basic data processing and modeling for time series involves "linear operators" applied to them. In this context it turns out to be a mathematical convenience (in the same way that calculus is a convenience for elementary science and technology in spite of the fact that the world is probably not continuous but rather discrete) to idealize most series as not finite in indexing, but rather doubly infinite in indexing. That is, instead of series (1) we imagine series

$$\boldsymbol{Y} = \begin{pmatrix} \vdots \\ y_{-2} \\ y_{-1} \\ y_0 \\ y_1 \\ y_2 \\ \vdots \end{pmatrix} \tag{2}$$

of which any real/observable series like (1) is a sub-vector. The vector (2) is formally an element of an infinite-dimensional Euclidean space, $\Re^\infty$ (while the observable vector (1) obviously belongs to n-space, $\Re^n$).

One conceptual advantage of formally considering infinite series like (2) is that operations often applied to time series can be thought of as operators/transformations/functions taking $\boldsymbol{Y}$ as an input and producing another

element of $\Re^\infty$ as an output. Versions of those same operations applied to finite vectors often lack meaning for at least some indices t.

So we consider **linear** operators/functions $\mathcal{L}$ on $\Re^\infty$ (or some appropriate subset of it). These have the property that for constants a and b,

$$\mathcal{L}(a\boldsymbol{Y} + b\boldsymbol{Z}) = a\mathcal{L}(\boldsymbol{Y}) + b\mathcal{L}(\boldsymbol{Z})$$

As a matter of notation, we will often not write the parentheses in $\mathcal{L}(\boldsymbol{Y})$, preferring instead to write the simpler $\mathcal{L}\boldsymbol{Y}$.

One particularly useful such operator is the **backshift operator** that essentially takes every entry in an input series $\boldsymbol{Y}$ and moves it ahead one index (slides the whole list of numbers in (2) "down" one slot). That is, using $\mathcal{B}$ to stand for this operator, if $\boldsymbol{Z} = \mathcal{B}\boldsymbol{Y}$, then

$$z_t = y_{t-1} \quad \forall t$$

We note that some authors sloppily write as if this operator somehow operates on individual values of a time series rather than on the whole series, writing the (nonsensical) expression $\mathcal{B}y_t = y_{t-1}$. (The expression $(\mathcal{B}\boldsymbol{Y})_t = y_{t-1}$ *would* make sense, since it says that the t entry of the infinite vector $\boldsymbol{Z} = \mathcal{B}\boldsymbol{Y}$ is y_{t-1}, the $t-1$ entry of $\boldsymbol{Y}$. But as it stands, the common notation is just confusing.)

The identity operator, $\mathcal{I}$, is more or less obviously defined by $\mathcal{I}\boldsymbol{Y} = \boldsymbol{Y}$. The composition of two (linear) operators, say $\mathcal{L}_1$ and $\mathcal{L}_2$, is what one gets upon following one by the other (this is ordinary mathematical composition). Employing parentheses for clarity

$$\mathcal{L}_1\mathcal{L}_2\boldsymbol{Y} \equiv \mathcal{L}_1 \circ \mathcal{L}_2(\boldsymbol{Y}) = \mathcal{L}_1(\mathcal{L}_2(\boldsymbol{Y})) \quad \forall \boldsymbol{Y}$$

A linear combination of (linear) operators is defined in the obvious way (in the same way that one defines linear combinations of finite-vector-valued functions of finite vectors) by

$$(a\mathcal{L}_1 + b\mathcal{L}_2)\boldsymbol{Y} \equiv a\mathcal{L}_1\boldsymbol{Y} + b\mathcal{L}_2\boldsymbol{Y} \quad \forall \boldsymbol{Y}$$

With these conventions, as long as one is careful not to reverse order of a "product" of operators (remembering that the "product" is really composition and composition doesn't obviously commute) one can do "ordinary algebra" on polynomials of operators (in the same way that one can do "ordinary algebra" on matrices as long is one is careful to not do violence to orders of multiplication of matrices).

1.1.1 Operators Based on the Backshift Operator

The facts above lead to a variety of interesting/useful operators, some based on differencing. The **first difference operator** is

$$\mathcal{D} = (\mathcal{I} - \mathcal{B})$$

If if $\boldsymbol{Z} = \mathcal{D}\boldsymbol{Y}$, then

$$z_t = y_t - y_{t-1} \quad \forall t$$

(or $(\mathcal{D}\boldsymbol{Y})_t = y_t - y_{t-1}$). The d**th difference operator** is defined by d compositions of the first difference operator

$$\mathcal{D}^d = \underbrace{\mathcal{D}\mathcal{D}\cdots\mathcal{D}}_{d \text{ factors}}$$

For example, it's easy to argue that if $\boldsymbol{Z} = \mathcal{D}^2\boldsymbol{Y}$ then

$$z_t = y_t - 2y_{t-1} + y_{t-2} \quad \forall t$$

(or $\left(\mathcal{D}^2\boldsymbol{Y}\right)_t = y_t - 2y_{t-1} + y_{t-2}$). In contexts where one expects some kind of "seasonality" in a time series at a spacing of s periods, a useful operator turns out to be a **seasonal difference operator of order** s

$$\mathcal{D}_s \equiv \mathcal{I} - \mathcal{B}^s$$

If $\boldsymbol{Z} = \mathcal{D}_s\boldsymbol{Y}$ then

$$z_t = y_t - y_{t-s} \quad \forall t$$

(or $(\mathcal{D}_s\boldsymbol{Y})_t = y_t - y_{t-s}$).

A generalization of these differencing operators that proves useful in time series modeling is that of **polynomial backshift operators**. That is, one might for example define an operator

$$\Phi(\mathcal{B}) = \mathcal{I} - \phi_1\mathcal{B}^1 - \phi_2\mathcal{B}^2 - \cdots - \phi_p\mathcal{B}^p$$

for real constants $\phi_1, \phi_2, \ldots, \phi_p$. If $\boldsymbol{Z} = \Phi(\mathcal{B})\boldsymbol{Y}$ then

$$z_t = y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} \quad \forall t$$

(or $(\Phi(\mathcal{B})\boldsymbol{Y})_t = y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p}$).

To take the polynomial backshift idea to its extreme, consider the expression

$$z_t = \sum_{s=-\infty}^{\infty} \psi_s y_{t-s} \quad \forall t \tag{3}$$

for some doubly infinite sequence of real constants $\ldots, \psi_{-2}, \psi_{-1}, \psi_0, \psi_1, \psi_2, \ldots$. Involving as it does infinite series, the expression (3) doesn't even make sense unless the ψ_t's and y_t's fit together well enough to guarantee convergence. Let us suppose that the weights ψ_t are absolutely summable, that is

$$\sum_{t=-\infty}^{\infty} |\psi_t| < \infty$$

Then the expression (3) at least makes sense if $\boldsymbol{Y}$ has entries bounded by some finite number (i.e. provided one cannot find a divergent sub-sequence of entries

of $\boldsymbol{Y}$). One can then define a linear operator, say $\mathcal{L}$, on the part of $\Re^\infty$ satisfying this boundedness condition using (3). That is, if $\boldsymbol{Z} = \mathcal{L}\boldsymbol{Y}$ the entries of it are given by (3)

$$(\mathcal{L}\boldsymbol{Y})_t = \sum_{s=-\infty}^{\infty} \psi_s y_{t-s} \quad \forall t \tag{4}$$

But this means that if we understand $\mathcal{B}^0$ to be $\mathcal{I}$, $\mathcal{B}^{-1}$ to mean a **forward shift operator** and $\mathcal{B}^{-k}$ to be the k-fold composition of this with itself (producing a forward shift by k places in the infinite vector being operated on)

$$\mathcal{L} = \sum_{s=-\infty}^{\infty} \psi_s \mathcal{B}^{t-s}$$

and this operator is a limit of polynomial backshift operators.

An operator defined by (4) is variously known as a **time-invariant linear filter**, a linear system, a linear transfer function, etc. It is apparently common to note that if $\boldsymbol{\Delta}_0$ is an element of $\Re^\infty$ with a 1 in the $t = 0$ position and 0's elsewhere,

$$(\mathcal{L}\boldsymbol{\Delta}_0)_t = \psi_t$$

so that $\mathcal{L}$ maps $\boldsymbol{\Delta}_0$ onto a vector that has its defining coefficients as elements. The "input" $\boldsymbol{\Delta}_0$ might then be called a "unit impulse (at time 0)" and the "output" vector of coefficients is often called the **impulse response (function)** of the filter. Further, when all ψ_s for $s < 0$ are 0, so that $(\mathcal{L}\boldsymbol{Y})_t$ depends only on those entries of $\boldsymbol{Y}$ with indices t or less, the linear filter $\mathcal{L}$ is sometimes called **causal** or **non-anticipatory**.

1.1.2 Linear Operators and Inverses

It is often useful to consider "undoing" a linear operation. This possibility is that of identifying an inverse (or at least a "left inverse") for a linear operator. In passing we noted above the obvious fact that the backshift operator has an inverse, the forward shift operator. That is, using $\mathcal{F}$ to stand for this operator, if $\boldsymbol{Z} = \mathcal{F}\boldsymbol{Y}$,

$$z_t = y_{t+1} \quad \forall t$$

(or $(\mathcal{F}\boldsymbol{Y})_t = y_{t+1}$). Then obviously $\mathcal{F}\mathcal{B} = \mathcal{I}$ and $\mathcal{F}$ is a left inverse for $\mathcal{B}$. It "undoes" the backshift operation. (It is also a right inverse for $\mathcal{B}$ and the backshift operator undoes it.)

Functions don't necessarily have inverses and linear operators don't always have inverses. (We're going to argue below that linear operators can be thought of as infinite-by-infinite matrices, and we all know that matrices don't have to have inverses.)

A very simple example that shows that even quite tame linear operators can fail to have inverses is the case of the first difference operator. That is, consider

the operator $\mathcal{D}$. If $\mathcal{D}\boldsymbol{Y} = \boldsymbol{Z}$ it is also the case that $\mathcal{D}(\boldsymbol{Y} + c\mathbf{1}) = \boldsymbol{Z}$ for c any scalar and $\mathbf{1}$ an infinite vector of 1's. This is easily seen since

$$z_t = y_t - y_{t-1} + c(1-1) = y_t - y_{t-1}$$

which means that (of course) there is no way to tell which vector has produced a given set of successive differences. The first difference operator is not invertible. (This is completely analogous to the fact that in calculus there are infinitely many functions that have a given derivative function, all of them differing by a constant. The first difference operator on time series is exactly analogous to the derivative operator on ordinary functions of a single real variable.)

As a more substantial example of a linear operator for which one *can* find an inverse, consider the operator $\mathcal{I} - \phi\mathcal{B}$ for a real value ϕ with $|\phi| < 1$ and a causal linear filter $\mathcal{L}$ defined at least for those time series with bounded entries by

$$(\mathcal{L}\boldsymbol{Y})_t = \sum_{s=0}^{\infty} \phi^s y_{t-s} \quad \forall t \tag{5}$$

(this is the case of filter (4) where $\psi_t = 0$ for $t < 0$ and $\psi_t = \phi^t$ otherwise). Then notice that the t entry of

$$\boldsymbol{Z} = \mathcal{L}(\mathcal{I} - \phi\mathcal{B})\boldsymbol{Y}$$

is

$$\begin{aligned} z_t &= \sum_{s=0}^{\infty} \phi^s (y_{t-s} - \phi y_{t-1-s}) \\ &= \sum_{s=0}^{\infty} \phi^s y_{t-s} - \sum_{s=0}^{\infty} \phi^{s+1} y_{t-(s+1)} \\ &= y_t \end{aligned}$$

(where the breaking apart of the first series for z_t into the difference of two is permissible because the boundedness of the entries of $\boldsymbol{Y}$ together with the fact that $|\phi| < 1$ means that both of the sums in the difference are absolutely convergent). So thinking about $\mathcal{L}$ and $\mathcal{I} - \phi\mathcal{B}$ as operators on those elements of $\Re^{\infty}$ with bounded entries, $\mathcal{L}$ functions as a left inverse for the operator $\mathcal{I} - \phi\mathcal{B}$. Notice that in light of (5) we might then want to write something like

$$(\mathcal{I} - \phi\mathcal{B})^{-1} = \sum_{s=0}^{\infty} \phi^s \mathcal{B}^s \tag{6}$$

As a practical matter one might for computational purposes truncate expression (6) at some sufficiently large upper limit to produce an approximate inverse for $\mathcal{I} - \phi\mathcal{B}$.

It is possible to in some cases generalize the previous example. Consider for real constants $\phi_1, \phi_2, \ldots, \phi_p$ the operator

$$\Phi(\mathcal{B}) = \sum_{j=1}^{p} \phi_j \mathcal{B}^j$$

and the polynomial backshift operator

$$\mathcal{I} - \Phi(\mathcal{B}) = \mathcal{I} - \phi_1\mathcal{B} - \phi_2\mathcal{B}^2 - \cdots - \phi_p\mathcal{B}^p$$

Then, define the finite series operator

$$\mathcal{L}_k = \sum_{s=0}^{k} (\Phi(\mathcal{B}))^s \tag{7}$$

and notice that

$$\begin{aligned}
\mathcal{L}_k(\mathcal{I} - \Phi(\mathcal{B})) &= \mathcal{L}_k - \left(\sum_{s=0}^{k} (\Phi(\mathcal{B}))^s\right)\Phi(\mathcal{B}) \\
&= \sum_{s=0}^{k} (\Phi(\mathcal{B}))^s - \sum_{s=1}^{k+1} (\Phi(\mathcal{B}))^s \\
&= \mathcal{I} - (\Phi(\mathcal{B}))^{k+1}
\end{aligned}$$

Now if the coefficients ϕ are such that with increasing k the operator $(\Phi(\mathcal{B}))^{k+1}$ becomes negligible, then we have that for large k the operator $\mathcal{L}_k$ is an approximate inverse for $\mathcal{I} - \Phi(\mathcal{B})$. We might in such cases write

$$(\mathcal{I} - \Phi(\mathcal{B}))^{-1} = \sum_{s=0}^{\infty} (\Phi(\mathcal{B}))^s$$

Conditions on the coefficients ϕ that will make this all work are conditions that guarantee that in some sense $\Phi(\mathcal{B})\boldsymbol{Y}$ is smaller than $\boldsymbol{Y}$. (Again consider the $p = 1$ case above and the condition that $|\phi| < 1$.)

Although $\mathcal{L}_k$ is a (conceptually simple) polynomial in the backshift operator (of order pk) there is no obvious easy way to find the associated coefficients or see limits for early ones. This particular exposition is then not so much a practical development as it is one intended to provide insight into structure of linear operators.

We proceed next to develop the connection between linear operators and matrices, and note in advance that the invertibility of a linear operator on time series is completely analogous to the invertibility of a finite square matrix.

1.1.3 Linear Operators and "Matrices"

In many respects, linear operators on $\Re^\infty$ amount to multiplication of infinitely long vectors by infinite-by-infinite matrices. I find this insight helpful and will demonstrate some of its use here. In order to make apparent where the "0 row" and "0 column" of an infinite-by-infinite matrix are and where the "0 position" of an infinitely long vector is, I will (in this discussion only) make bold face the values in those rows, columns, and positions. Rows of the matrices should be thought of as indexed from $-\infty$ to ∞ top to bottom and columns indexed from $-\infty$ to ∞ left to right.

First notice that one might conceive of the backshift operation in matrix multiplication terms as

$$\mathcal{B}\boldsymbol{Y} = \begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & \mathbf{0} & 0 & 0 & \cdots \\ \cdots & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \cdots & 0 & 0 & \mathbf{1} & 0 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} \vdots \\ y_{-2} \\ y_{-1} \\ \boldsymbol{y}_0 \\ y_1 \\ y_2 \\ \vdots \end{pmatrix}$$

(as always, one lines up rows of the matrix alongside the vector and multiplies values next to each other and sums those products, here being careful to get the element in the 0 column positioned next to the $t = 0$ entry of the vector). In contrast to this, the forward shift operation might be represented as

$$\mathcal{F}\boldsymbol{Y} = \begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 0 & 0 & \mathbf{1} & 0 & 0 & \cdots \\ \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \cdots \\ \cdots & 0 & 0 & \mathbf{0} & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} \vdots \\ y_{-2} \\ y_{-1} \\ \boldsymbol{y}_0 \\ y_1 \\ y_2 \\ \vdots \end{pmatrix}$$

and, of course, the identity operation can be represented as

$$\mathcal{I}\boldsymbol{Y} = \begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 0 & 1 & \mathbf{0} & 0 & 0 & 0 & \cdots \\ \cdots & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \cdots & 0 & 0 & \mathbf{0} & 1 & 0 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} \vdots \\ y_{-2} \\ y_{-1} \\ \boldsymbol{y}_0 \\ y_1 \\ y_2 \\ \vdots \end{pmatrix}$$

The operator $\mathcal{I} - \phi\mathcal{B}$ might be represented by the matrix

$$\begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & -\phi & 1 & \mathbf{0} & 0 & 0 & \cdots \\ \cdots & \mathbf{0} & \boldsymbol{-\phi} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots \\ \cdots & 0 & 0 & \boldsymbol{-\phi} & 1 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

and its inverse operator might be represented by

$$\begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & \phi & 1 & \mathbf{0} & 0 & 0 & 0 & \cdots \\ \cdots & \boldsymbol{\phi^2} & \boldsymbol{\phi} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \cdots & \phi^3 & \phi^2 & \boldsymbol{\phi} & 1 & 0 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

In fact, a general time-invariant linear filter might be represented by

$$\begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & \psi_1 & \psi_0 & \boldsymbol{\psi_{-1}} & \psi_{-2} & \psi_{-3} & \cdots \\ \cdots & \boldsymbol{\psi_2} & \boldsymbol{\psi_1} & \boldsymbol{\psi_0} & \boldsymbol{\psi_{-1}} & \boldsymbol{\psi_{-2}} & \cdots \\ \cdots & \psi_3 & \psi_2 & \boldsymbol{\psi_1} & \psi_0 & \psi_{-1} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

and the 0 column (or reversed 0 row) in the matrix gives the impulse response function for the filter. Note that for two time-invariant linear filters, say $\mathcal{A}$ and $\mathcal{B}$, represented by say matrices

$$\begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & \alpha_1 & \alpha_0 & \boldsymbol{\alpha_{-1}} & \alpha_{-2} & \alpha_{-3} & \cdots \\ \cdots & \boldsymbol{\alpha_2} & \boldsymbol{\alpha_1} & \boldsymbol{\alpha_0} & \boldsymbol{\alpha_{-1}} & \boldsymbol{\alpha_{-2}} & \cdots \\ \cdots & \alpha_3 & \alpha_2 & \boldsymbol{\alpha_1} & \alpha_0 & \alpha_{-1} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & \beta_1 & \beta_0 & \boldsymbol{\beta_{-1}} & \beta_{-2} & \beta_{-3} & \cdots \\ \cdots & \boldsymbol{\beta_2} & \boldsymbol{\beta_1} & \boldsymbol{\beta_0} & \boldsymbol{\beta_{-1}} & \boldsymbol{\beta_{-2}} & \cdots \\ \cdots & \beta_3 & \beta_2 & \boldsymbol{\beta_1} & \beta_0 & \beta_{-1} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

the fact that $\mathcal{B}\boldsymbol{Y}$ can be represented by an infinite matrix multiplication (as can $\mathcal{A}\boldsymbol{Z}$) means that the composition composition linear operator $\mathcal{L} = \mathcal{A}\mathcal{B}$ can be represented by the product of the above matrices. The matrix representing $\mathcal{L}$ has in its (t, s) position

$$l_{t,s} = \sum_{j=-\infty}^{\infty} \alpha_{t-j}\beta_{-s+j} = \sum_{l=-\infty}^{\infty} \alpha_l \beta_{(t-s)-l}$$

That is, the product of the two infinite-by-infinite matrices representing $\mathcal{L} = \mathcal{A}\mathcal{B}$ is

$$\begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & \sum_{l=-\infty}^{\infty} \alpha_{-l}\beta_{l+1} & \sum_{l=-\infty}^{\infty} \alpha_{-l}\beta_l & \sum_{l=-\infty}^{\infty} \boldsymbol{\alpha_{-l}\beta_{l-1}} & \sum_{l=-\infty}^{\infty} \alpha_{-l}\beta_{l-2} & \sum_{l=-\infty}^{\infty} \alpha_{-l}\beta_{l-3} & \cdots \\ \cdots & \sum_{l=-\infty}^{\infty} \boldsymbol{\alpha_{-l}\beta_{l+2}} & \sum_{l=-\infty}^{\infty} \boldsymbol{\alpha_{-l}\beta_{l+1}} & \sum_{l=-\infty}^{\infty} \boldsymbol{\alpha_{-l}\beta_l} & \sum_{l=-\infty}^{\infty} \boldsymbol{\alpha_{-l}\beta_{l-1}} & \sum_{l=-\infty}^{\infty} \boldsymbol{\alpha_{-l}\beta_{l-2}} & \cdots \\ \cdots & \sum_{l=-\infty}^{\infty} \alpha_{-l}\beta_{l+3} & \sum_{l=-\infty}^{\infty} \alpha_{-l}\beta_{l+2} & \sum_{l=-\infty}^{\infty} \boldsymbol{\alpha_{-l}\beta_{l+1}} & \sum_{l=-\infty}^{\infty} \alpha_{-l}\beta_l & \sum_{l=-\infty}^{\infty} \alpha_{-l}\beta_{l-1} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}$$

It's then the case that $\mathcal{L}$ is itself a time-invariant linear filter with

$$\psi_s = l_{0,-s} = \sum_{l=-\infty}^{\infty} \alpha_l \beta_{s-l}$$

(representing the convolution of the two impulse response functions) and absolute summability of the α's and β's guarantees that of the ψ's.

1.2 Initial Probability Modeling Ideas for Time Series

In one sense, there is nothing "new" in probability modeling for time series beyond what is in a basic probability course. It is just multivariate probability modeling. But there are some complicated things special to honoring the significance of time ordering of the variables and dealing with the probability implications of the "infinite sequence of variables" idealization (that is so convenient because linear operators are such nice tools for time series modeling and data analysis). Before getting seriously into the details of modeling, and inference and prediction based on the modeling, it seems potentially useful to give a "50,000 ft" view of the landscape.

We first recall several basics of multivariate distributions/probability modeling. For an n-dimensional random vector $\boldsymbol{Y}$ (we're effectively now talking about giving the series in (1) a probability distribution) with mean vector and covariance matrix respectively

$$\boldsymbol{\mu} = \mathrm{E}\boldsymbol{Y} = \begin{pmatrix} \mathrm{E}y_1 \\ \mathrm{E}y_2 \\ \vdots \\ \mathrm{E}y_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \mathrm{Var}\boldsymbol{Y} = \left(\mathrm{Cov}\left(y_i, y_j\right)\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

and an $n \times n$ matrix $\boldsymbol{M}$, the random vector $\boldsymbol{Z} = \boldsymbol{M}\boldsymbol{Y}$ has mean and covariance respectively

$$\mathrm{E}\boldsymbol{Z} = \mathrm{E}\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{\mu} \quad \text{and} \quad \mathrm{Var}\boldsymbol{Z} = \mathrm{Var}\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}'$$

Focusing attention on only some of the entries of $\boldsymbol{Y}$, we find ourselves talking about a (joint, because more than one coordinate may be involved) marginal distribution of the full model for $\boldsymbol{Y}$. In terms of simply means and variances/covariances, the mean vector for a sub-vector of $\boldsymbol{Y}$ is simply the corresponding sub-vector of $\boldsymbol{\mu}$, and the covariance matrix is obtained by deleting from $\boldsymbol{\Sigma}$ all rows and columns corresponding to the elements of $\boldsymbol{Y}$ not of interest.

We mention these facts for n-dimensional distributions because a version of them is true regarding models for the infinite-dimensional case as well. If we can successfully define a distribution for the series (2) then linear operations on it have means and variances/covariances that are easily understood from those of the original model, and realizable/observable/finite parts (1) of the series (2) have models (and means and covariances) that are just read directly as marginals from the theoretically infinite-dimensional model.

Much of statistical analysis conforms to a basic conceptualization that

$$\textit{what is observable} = \textit{signal} + \textit{noise}$$

where the "signal" is often a mean that can be a parametric function of one or more explanatory variables and the "noise" is ideally fairly small and "random." Time series models don't much depart from this paradigm except that because of the relevance of time order, there can be much more potentially interesting and useful structure attributed to the "noise." If departures from the norm/signal in a time series tend to be correlated, then prediction of a "next" observation can take account not only of signal/trend but also the nature of that correlation.

A basic kind of time series modeling then begins with a probability model for an infinite vector of random variables

$$\boldsymbol{\epsilon} = \begin{pmatrix} \vdots \\ \epsilon_{-2} \\ \epsilon_{-1} \\ \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \vdots \end{pmatrix}$$

that has $\mathrm{E}\boldsymbol{\epsilon} = \mathbf{0}$ and $\mathrm{Var}\boldsymbol{\epsilon} = \sigma^2\boldsymbol{I}$. These assumptions on $\boldsymbol{\epsilon}$ about means and variances are usually called **white noise** assumptions. (A model assumption that the ϵ's are iid/independent random draws from some distribution with mean 0 and standard deviation σ implies the white noise conditions but isn't necessary to produce them.)

A way to move from uncorrelated noise to models with correlation between successive observations is to consider

$$\mathcal{N}\boldsymbol{\epsilon}$$

for some linear operator $\mathcal{N}$ that "works" (is mathematically convenient and produces useful/appealing kinds of correlations). One might expect that if $\mathcal{N}$ can be represented by an infinite-by-infinite matrix $\boldsymbol{N}$ and $\mathcal{N}\boldsymbol{\epsilon}$ makes sense with probability 1 (the white noise model for $\boldsymbol{\epsilon}$ can't put positive probability on the set of elements of $\Re^\infty$ for which the expression is meaningless)

$$\mathrm{Var}\mathcal{N}\boldsymbol{\epsilon} = \boldsymbol{N}\sigma^2\boldsymbol{I}\boldsymbol{N}' = \sigma^2\boldsymbol{N}\boldsymbol{N}'$$

(provided I can convince myself that $\boldsymbol{N}\boldsymbol{N}'$ makes sense).

Then with

$$\boldsymbol{X}_i = \text{an } i\text{th "exogenous" or predictor series}$$

and

$$\mathcal{D}^* = \text{some appropriate differencing operator}$$

and a set of linear operators $\mathcal{L}, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k$ (that could, for example, be polynomial backshift operators) I might model as

$$\mathcal{L}\mathcal{D}^*\boldsymbol{Y} = \sum_{i=1}^{k} \mathcal{L}_i \mathcal{D}^* \boldsymbol{X}_i + \mathcal{N}\boldsymbol{\epsilon} \tag{8}$$

(The differencing of the response series $\mathcal{D}^*\boldsymbol{Y}$ and the corresponding differencing of the predictor series $\mathcal{D}^*\boldsymbol{X}_i$ are typically done to remove trend and seasonality from the raw series. There is usually no non-zero mean here because of the differencing and the fact that including it would thereby imply explosive large n behavior of the original $\boldsymbol{Y}$.) If we write $\boldsymbol{Y}^* = \mathcal{D}^*\boldsymbol{Y}$ and $\boldsymbol{X}_i^* = \mathcal{D}^*\boldsymbol{X}_i$ this model can be written as

$$\mathcal{L}\boldsymbol{Y}^* = \sum_{i=1}^{k} \mathcal{L}_i \boldsymbol{X}_i^* + \mathcal{N}\boldsymbol{\epsilon}$$

and if $\mathcal{L}$ has an inverse perhaps this boils down to

$$\boldsymbol{Y}^* = \sum_{i=1}^{k} \mathcal{L}^{-1}\mathcal{L}_i \boldsymbol{X}_i^* + \mathcal{L}^{-1}\mathcal{N}\boldsymbol{\epsilon} \tag{9}$$

Model (9) begins to look a lot like a regression model where a transformed response is assumed to have a mean that is a linear form involving transformed inputs, and "errors" that have mean $\mathbf{0}$ and a covariance matrix $\sigma^2 \boldsymbol{L}^{-1}\boldsymbol{N}\boldsymbol{N}'\left(\boldsymbol{L}^{-1}\right)'$ (where $\boldsymbol{L}^{-1}$ is the matrix representing $\mathcal{L}^{-1}$). Now all of $\mathcal{N}, \mathcal{L}, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k$ are typically hiding parameters (like the coefficients in backshift polynomials) so the whole business of fitting a model like (9) by estimating those parameters and then making corresponding forecasts/predictions is not so trivial. But at least conceptually, this kind of form should now not seem all that surprising or mysterious ... and all else is just detail ... ;+}.

The class of models that are of form (8) with $\mathcal{N}$ and $\mathcal{L}$ polynomial backshift operators is the "ARIMAX" class (autoregressive integrated moving-average exogenous variables models). The "I" refers to the fact that differencing has been done and "integration"/summing is needed to get back to the original response, and the "X" refers to the presence of the predictor series. The special case without differencing or predictors

$$\mathcal{L}\boldsymbol{Y} = \mathcal{N}\boldsymbol{\epsilon}$$

is the famous ARIMA class, and we remark (as an issue to be revisited more carefully soon) that provided $\mathcal{L}^{-1}$ exists, in this class we can hope that

$$\boldsymbol{Y} = \mathcal{L}^{-1}\mathcal{N}\boldsymbol{\epsilon}$$

and the model for $\boldsymbol{Y}$ might boil down to that of a time-invariant linear filter applied to white noise.

2 Stationarity and Linear Processes

2.1 Basics of Stationary and Linear Processes

We begin in earnest to consider distributions/probability models for time series. In any statistical modeling and inference, there must be some things that don't change across the data set, constituting structure that is to be discovered and quantified. In time series modeling the notions of unchangeableness are given precise meanings and called "stationarity."

A distribution for a time series $\boldsymbol{Y}$ is **strictly stationary** ($\boldsymbol{Y}$ is strictly stationary) if $\mathcal{B}^k \boldsymbol{Y}$ has the same distribution as $\boldsymbol{Y}$ for every integer k (positive or negative, understanding that $\mathcal{B}^{-1} = \mathcal{F}$). This, of course, implies that for every k and $l \geq 1$ the vectors

$$(y_1, \ldots, y_l) \text{ and } (y_{k+1}, \ldots, y_{k+l})$$

have the same (joint) distributions. This is a strong mathematical condition and more than is needed to support most standard time series analysis. Instead, the next concept typically suffices.

$\boldsymbol{Y}$ is said to have a **second order** (or wide sense or weakly) **stationary distribution** if every $\mathrm{E}y_t^2 < \infty$ and

$$\mathrm{E}y_t = \mu \text{ (some constant not depending upon } t)$$

and $\mathrm{Cov}(y_t, y_{t+s})$ is independent of t, in which case we can write

$$\gamma(s) = \mathrm{Cov}(y_t, y_{t+s})$$

and call $\gamma(s)$ the **autocovariance function** for the process. Note that $\gamma(0) = \mathrm{Var}y_t$ for all t, $\gamma(-s) = \gamma(s)$, and that the ratio

$$\rho(s) \equiv \frac{\gamma(s)}{\gamma(0)}$$

provides the **autocorrelation function** for the process.

If $\boldsymbol{\epsilon}$ is white noise and $\mathcal{L}$ is a time-invariant linear operator with $\sum_{t=-\infty}^{\infty} |\psi_t| < \infty$ then it's a (not necessarily immediately obvious) fact that the output of

$$\mathcal{L}\boldsymbol{\epsilon}$$

is well-defined with probability 1. (The probability with which any one of the series defining the entries of $\boldsymbol{Y} = \mathcal{L}\boldsymbol{\epsilon}$ fails to converge is 0. See the Brockwell and Davis "Methods" book (henceforth BDM) page 51.) In fact, each $\mathrm{E}y_t^2 < \infty$,

the distribution of $\boldsymbol{Y}$ is second order stationary, and

$$\begin{aligned}
\gamma(s) &= \operatorname{Cov}(y_t, y_{t+s}) \\
&= \operatorname{Cov}\left(\sum_{i=-\infty}^{\infty} \psi_i \epsilon_{t-i}, \sum_{i=-\infty}^{\infty} \psi_i \epsilon_{t+s-i}\right) \\
&= \operatorname{Cov}\left(\sum_{j=-\infty}^{\infty} \psi_{t-j} \epsilon_j, \sum_{j=-\infty}^{\infty} \psi_{t+s-j} \epsilon_j\right) \\
&= \sigma^2 \sum_{j=-\infty}^{\infty} \psi_{t-j} \psi_{s+t-j} \\
&= \sigma^2 \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+s} \qquad (10)
\end{aligned}$$

In particular

$$\gamma(0) = \sigma^2 \sum_{i=-\infty}^{\infty} \psi_i^2$$

In this context, it is common to call $\mathcal{L}\boldsymbol{\epsilon}$ a **linear process**.

The class of linear process models is quite rich. Wold's decomposition (BDM pages 51 and 77+) says that every second order process is either a linear process or differs from one only by a "deterministic" series. (See BDM for the technical meaning of "deterministic" in this context.) Further, the fact that a time invariant linear filter operating on white noise produces a second order stationary process generalizes beyond white noise to wide sense stationary processes. That is, Proposition 2.2.1 of BDM states the following.

Proposition 1 *If $\boldsymbol{Y}$ is wide sense stationary with mean $\mathbf{0}$ and autocovariance function $\gamma_{\boldsymbol{Y}}$ and $\mathcal{L}$ is a time invariant linear filter with $\sum_{t=-\infty}^{\infty} |\psi_t| < \infty$, then*

$$\mathcal{L}\boldsymbol{Y}$$

is well-defined with probability 1, $Ey_t^2 < \infty$ for each t, and $\mathcal{L}\boldsymbol{Y}$ is wide sense stationary. Further, $\mathcal{L}\boldsymbol{Y}$ has autocovariance function

$$\gamma_{\mathcal{L}\boldsymbol{Y}}(s) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_{\boldsymbol{Y}}(s + (k-j))$$

(The form for the autocovariance is, of course, what follows from the infinite-by-infinite matrix calculation of a covariance matrix via $\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}'$.)

2.2 MA(q) and AR(1) Models

The **moving average processes of order** q (MA(q) processes) are very important elementary instances of linear processes. That is, for

$$\Theta(\mathcal{B}) = \mathcal{I} + \sum_{j=1}^{q} \theta_j \mathcal{B}^j$$

and $\boldsymbol{\epsilon}$ white noise,

$$\boldsymbol{Y} = \Theta(\mathcal{B})\,\boldsymbol{\epsilon}$$

is a MA(q) process. Alternative notation here is that

$$y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}\ \forall t$$

and (with the convention that $\theta_0 = 1$) it is possible to argue that the autocovariance function for $\boldsymbol{Y}$ is

$$\gamma(s) = \begin{cases} \sigma^2\left(\theta_s + \theta_1\theta_{s+1} + \cdots + \theta_{q-s}\theta_q\right) & \text{if } |s| \le q \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

(Unsurprisingly, the covariances at lags bigger than q are 0.)

Consider next a model specified using the operator

$$\Phi(\mathcal{B}) = \mathcal{I} - \phi\mathcal{B}$$

for $|\phi| < 1$. A model for $\boldsymbol{Y}$ satisfying

$$y_t = \phi y_{t-1} + \epsilon_t \quad \forall t \tag{12}$$

satisfies

$$\Phi(\mathcal{B})\,\boldsymbol{Y} = \boldsymbol{\epsilon} \tag{13}$$

and might be called an autoregressive model of order 1 (an AR(1) model). Now we have seen that where $|\phi| < 1$

$$\mathcal{L} = \sum_{j=0}^{\infty}\phi^j\mathcal{B}^j \tag{14}$$

is a time-invariant linear operator with

$$\psi_j = \begin{cases} \phi^j & \text{for } j \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

(and thus absolutely summable coefficients) that is an inverse for $\Phi(\mathcal{B})$. So in this context,

$$\boldsymbol{Y} = \mathcal{L}\boldsymbol{\epsilon}$$

is a linear process that solves the equation (13) with probability 1. (BDM argue that in fact it is the only stationary solution to the equation, so that its properties are implied by the equation.) Such a $\boldsymbol{Y}$ has

$$y_t = \sum_{j=0}^{\infty}\phi^j\epsilon_{t-j}$$

and has autocovariance function

$$\gamma(s) = \sigma^2\sum_{j=0}^{\infty}\phi^j\phi^{j+s} = \sigma^2\frac{\phi^{|s|}}{1-\phi^2}$$

and autocorrelation function

$$\rho(s) = \phi^{|s|} \tag{15}$$

Next consider the version of the equation (12) where $|\phi| > 1$ as a possible device for specifying a second order stationary model for $\boldsymbol{Y}$. The development above falls apart in this case because $\mathcal{I} - \phi\mathcal{B}$ has no inverse. But there is this. Rewrite equation (12) as

$$y_{t-1} = \frac{1}{\phi} y_t - \frac{1}{\phi}\epsilon_t \quad \forall t \tag{16}$$

For $\mathcal{F}$ the forward shift operator, in operator notation relationship (16) is

$$\boldsymbol{Y} = \frac{1}{\phi}\mathcal{F}\boldsymbol{Y} - \frac{1}{\phi}\mathcal{F}\boldsymbol{\epsilon}$$

or

$$\left(\mathcal{I} - \frac{1}{\phi}\mathcal{F}\right)\boldsymbol{Y} = -\frac{1}{\phi}\mathcal{F}\boldsymbol{\epsilon} \tag{17}$$

Now

$$\boldsymbol{\epsilon}^* \equiv -\frac{1}{\phi}\mathcal{F}\boldsymbol{\epsilon}$$

is white noise with variance σ^2/ϕ^2 and since $\left|\phi^{-1}\right| < 1$ essentially the same arguments applied above to identify an inverse for $\mathcal{I} - \phi\mathcal{B}$ in $|\phi| < 1$ cases show that $\left(\mathcal{I} - \frac{1}{\phi}\mathcal{F}\right)$ has an inverse

$$\mathcal{L} = \sum_{j=0}^{\infty} \phi^{-j}\mathcal{F}^j$$

that is a time-invariant linear operator with

$$\psi_j = \begin{cases} \phi^j & \text{for } j \le 0 \\ 0 & \text{otherwise} \end{cases}$$

(and thus absolutely summable coefficients). Thus in this context,

$$\boldsymbol{Y} = \mathcal{L}\boldsymbol{\epsilon}^*$$

is a linear process that solves the equation (17) with probability 1, and it is the only stationary solution to the equation. Notice that

$$\begin{aligned} \mathcal{L}\boldsymbol{\epsilon}^* &= -\frac{1}{\phi}\left(\sum_{j=0}^{\infty}\phi^{-j}\mathcal{F}^j\right)\mathcal{F}\boldsymbol{\epsilon} \\ &= -\left(\sum_{j=1}^{\infty}\phi^{-j}\mathcal{F}^j\right)\boldsymbol{\epsilon} \end{aligned}$$

so that with probability 1

$$y_t = -\sum_{j=1}^{\infty} \phi^{-j} \epsilon_{t+j} \quad \forall t \tag{18}$$

and $\boldsymbol{Y}$ has the representation of a linear filter with coefficients

$$\psi_j = \begin{cases} -\phi^j & \text{for } j \leq -1 \\ 0 & \text{otherwise} \end{cases}$$

applied to $\boldsymbol{\epsilon}$. This filter has less-than-intuitively-appealing property of failing to be causal. So as a means of specifying a second order stationary time series model, the equation (12) where $|\phi| > 1$ leaves something to be desired.

As a way out of this unpleasantness, consider the autocovariance function implied by expression (18). According to expression (10) this is

$$\gamma(s) = \frac{\sigma^2}{\phi^2} \left(\frac{1}{\phi}\right)^{|s|} \frac{1}{1 - \left(\frac{1}{\phi}\right)^2}$$

producing autocorrelation function

$$\rho(s) = \left(\frac{1}{\phi}\right)^{|s|} \tag{19}$$

Comparing this expression (19) to the autocorrelation function for the $|\phi| < 1$ version of an AR(1) model in display (15), we see that the parameters $|\phi| > 1$ generate the same set of correlation structures as do the parameters $|\phi| < 1$. This is a problem of lack of **identifiability**. All we'll really ever be able to learn from data from a stationary model are a mean, a marginal variance, and a correlation structure, and the $|\phi| > 1$ and $|\phi| < 1$ cases generate exactly the same sets of 2nd order summaries. One set is therefore redundant, and the common way out of this problem is to simply say that the $|\phi| < 1$ set is mathematically more pleasing and so we'll take the AR(1) parameter space to exclude values $|\phi| > 1$. (Time series authors seem to like using language like "We'll restrict attention to models with $|\phi| < 1$." I find that language confusing. There is no exclusion of possible second order moment structures. There is simply the realization that a given AR(1) structure comes from two different ϕ's if all real values of the parameter are allowed, and a decision is then made to reduce the parameter set by picking the possibility that has the most appealing mathematics.) BDM refer to the $|\phi| < 1$ restriction of parameters as the choice to consider only **causal** (linear) **processes**. This language is (from my perspective) substantially better than the more common language (probably traceable to Box and Jenkins) that terms the choice one of restriction to "stationary" processes. After all, we've just argued clearly that there are stationary solutions to basic AR(1) model equation even in the event that one considers $|\phi| > 1$.

For completeness sake, it might perhaps be noted that if $\phi = \pm 1$ there is no stationary model for which equation (12) makes sense.

Motivated by consideration of restriction of the AR(1) parameter space, let us briefly revisit the MA(1) model, and in particular consider the autocorrelation function that follows from autocovariance (11). That is

$$\rho(s) = \begin{cases} 1 & \text{if } s = 0 \\ \dfrac{\theta_1}{1+\theta_1^2} & \text{if } |s| = 1 \\ 0 & \text{if } |s| > 1 \end{cases} \tag{20}$$

Notice now that the choices $\theta_1 = c$ and $\theta_1 = 1/c$ for a non-zero real number c produce exactly the same autocorrelation function. That is, there is an indentifiability issue for MA(1) models exactly analogous to that for the AR(1) models. The set of MA(1) parameters θ with $|\theta| < 1$ generates the same set of correlation structures as does the set of MA(1) parameters θ with $|\theta| > 1$. So if one wants an unambiguous representation of MA(1) autocorrelation functions, some choice needs to be made. It is common to make the choice to exclude MA(1) parameters θ with $|\theta| > 1$. It seems common to then call the MA(1) models with parameter $|\theta| < 1$ **invertible**. I suppose that is most fundamentally because the operator $\Theta(\mathcal{B}) = \mathcal{I} + \theta\mathcal{B}$ used in the basic MA(1) equation $\boldsymbol{Y} = \Theta(\mathcal{B})\boldsymbol{\epsilon}$ is invertible (has an inverse) when $|\theta| < 1$. Some authors talk about the fact that when $\Theta(\mathcal{B}) = \mathcal{I} + \theta\mathcal{B}$ is invertible, one then has an "infinite series in the elements of $\boldsymbol{Y}$" (or infinite regression) representation for the elements of $\boldsymbol{\epsilon}$. While true, it's not obvious to me why this latter fact is of much interest.

2.3 ARMA(1,1) Models

A natural extension of both the MA(1) and AR(1) modeling ideas is the possibility of using the (autoregressive moving average/ ARMA$(1,1)$) equation

$$(\mathcal{I} - \phi\mathcal{B})\boldsymbol{Y} = (\mathcal{I} + \theta\mathcal{B})\boldsymbol{\epsilon} \tag{21}$$

for $\boldsymbol{\epsilon}$ white noise to potentially specify a second order stationary model for time series $\boldsymbol{Y}$. In other symbols, this is

$$y_t - \phi y_{t-1} = \epsilon_t + \theta\epsilon_{t-1} \quad \forall t$$

From what was just said about MA(1) models, it is clear that every autocovariance structure available for $(\mathcal{I} + \theta\mathcal{B})\boldsymbol{\epsilon}$ on the right of equation (21) using $|\theta| > 1$ is also available with a choice of $|\theta| < 1$ and that in order to avoid lack of identifiability one needs to restrict the parameter space for θ. It is thus standard to agree to represent the possible covariance structures for $(\mathcal{I} + \theta\mathcal{B})\boldsymbol{\epsilon}$ without using parameters $|\theta| > 1$. Using the same language as was introduced in the MA(1) context, we choose an "invertible" representation of ARMA$(1,1)$ models.

Next, as in the AR(1) discussion, when $|\phi| < 1$ the operator $\mathcal{L}$ defined in display (14) is an inverse for $\mathcal{I} - \phi\mathcal{B}$ so that

$$\boldsymbol{Y} = \mathcal{L}\left(\mathcal{I} - \phi\mathcal{B}\right)\boldsymbol{Y} = \mathcal{L}\left(\mathcal{I} + \theta\mathcal{B}\right)\boldsymbol{\epsilon}$$

is stationary (applying Proposition 1 to the stationary $\left(\mathcal{I} + \theta\mathcal{B}\right)\boldsymbol{\epsilon}$) and solves the ARMA$(1,1,)$ equation with probability 1. The operator on $\boldsymbol{\epsilon}$ is

$$\begin{aligned}
\mathcal{L}\left(\mathcal{I} + \theta\mathcal{B}\right) &= \left(\sum_{j=0}^{\infty}\phi^j\mathcal{B}^j\right)\left(\mathcal{I} + \theta\mathcal{B}\right) \\
&= \sum_{j=0}^{\infty}\phi^j\mathcal{B}^j + \theta\sum_{j=0}^{\infty}\phi^j\mathcal{B}^{j+1} \\
&= \mathcal{I} + \left(\phi + \theta\right)\sum_{j=1}^{\infty}\phi^{j-1}\mathcal{B}^j
\end{aligned}$$

so the ψ_j's for this time-invariant linear filter are 0 for $j < 0$, $\psi_0 = 1$, and $\psi_j = \left(\phi + \theta\right)\phi^{j-1}$ for $j > 1$. The autocovariance function for $\mathcal{L}\left(\mathcal{I} + \theta\mathcal{B}\right)\boldsymbol{\epsilon}$ implied by these is derived many places (including the original book of Box and Jenkins) and has the form

$$\gamma\left(s\right) = \begin{cases} \dfrac{1 + \theta^2 + 2\phi\theta}{1 - \phi^2}\sigma^2 & \text{for } s = 0 \\ \dfrac{\left(1 + \phi\theta\right)\left(\phi + \theta\right)}{1 - \phi^2}\sigma^2 & \text{for } |s| = 1 \\ \phi\gamma\left(|s| - 1\right) & \text{for } |s| > 1 \end{cases}$$

which in turn produces the autocorrelation function

$$\rho\left(s\right) = \begin{cases} 1 & \text{for } s = 0 \\ \dfrac{\left(1 + \phi\theta\right)\left(\phi + \theta\right)}{1 + \theta^2 + 2\phi\theta} & \text{for } |s| = 1 \\ \phi^{|s|-1}\rho\left(1\right) & \text{for } |s| > 1 \end{cases}$$

We might suspect that possible representations of ARMA$(1,1)$ autocovariance structures in terms of AR coefficient $|\phi| > 1$ are redundant once one has considered $|\phi| < 1$ cases. The following is an argument to that effect. Using the same logic as was applied in the AR(1) discussion, for $|\phi| > 1$, since $\left(\mathcal{I} + \theta\mathcal{B}\right)\boldsymbol{\epsilon}$ is stationary, for the time invariant linear filter

$$\mathcal{L} = -\left(\sum_{j=1}^{\infty}\phi^{-j}\mathcal{F}^j\right)$$

the equation

$$\left(\mathcal{I} - \phi\mathcal{B}\right)\boldsymbol{Y} = \left(\mathcal{I} + \theta\mathcal{B}\right)\boldsymbol{\epsilon}$$

has a unique stationary solution solution that with probability 1 can be represented as

$$\mathcal{L}\left(\mathcal{I}+\theta\mathcal{B}\right)\boldsymbol{\epsilon}$$

That is, in the case where $|\phi|<1$ the coefficients of the time-invariant linear operator $\mathcal{L}$ applied to $\boldsymbol{Z}=\left(\mathcal{I}+\theta\mathcal{B}\right)\boldsymbol{\epsilon}$ to produce a stationary solution for the ARMA$(1,1)$ equation are

$$\psi_j=\left\{\begin{array}{cc}\phi^j & j\geq 0\\ 0 & j<0\end{array}\right.$$

and in the $|\phi|>1$ case they are

$$\psi_j=\left\{\begin{array}{cc}0 & j>-1\\ -\phi^j & j\leq -1\end{array}\right.$$

Then applying the form for the autocovariance function of a linear filter applied to a stationary process given in Proposition 1, for the $|\phi|<1$ case, the autocovariance function for $\mathcal{L}\left(\mathcal{I}+\theta\mathcal{B}\right)\boldsymbol{\epsilon}$ is

$$\gamma^{*}_{\mathcal{L}\boldsymbol{Z}}\left(s\right)=\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\phi^j\phi^k\gamma_{\boldsymbol{Z}}\left(s+(k-j)\right)=\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\phi^{j+k}\gamma_{\boldsymbol{Z}}\left(s+k-j\right)$$

For the $|\phi|>1$ case, the autocovariance function for $\mathcal{L}\left(\mathcal{I}+\theta\mathcal{B}\right)\boldsymbol{\epsilon}$ is

$$\begin{aligned}\gamma^{**}_{\mathcal{L}\boldsymbol{Z}}\left(s\right) &= \sum_{j\leq -1}\sum_{k\leq -1}\left(-\phi^j\right)\left(-\phi^k\right)\gamma_{\boldsymbol{Z}}\left(s+k-j\right)\\ &= \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\left(\frac{1}{\phi}\right)^j\left(\frac{1}{\phi}\right)^k\gamma_{\boldsymbol{Z}}\left(s+j-k\right)\\ &= \left(\frac{1}{\phi}\right)^2\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\left(\frac{1}{\phi}\right)^{j-1}\left(\frac{1}{\phi}\right)^{k-1}\gamma_{\boldsymbol{Z}}\left(s+(j-1)-(k-1)\right)\\ &= \left(\frac{1}{\phi}\right)^2\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\left(\frac{1}{\phi}\right)^{j+k}\gamma_{\boldsymbol{Z}}\left(s+k-j\right)\end{aligned}$$

So, using $\phi=c$ with $|c|<1$ in the first case and the corresponding $\phi=1/c$ in the second produces

$$\gamma^{**}_{\mathcal{L}\boldsymbol{Z}}\left(s\right)=c^2\gamma^{*}_{\mathcal{L}\boldsymbol{Z}}\left(s\right)$$

The two autocovariance functions differ only by a constant multiplier and thus produce the same autocorrelation functions. That is, considering ARMA$(1,1)$ models with AR parameter $|\phi|>1$ only reproduces the set of correlation functions available using $|\phi|<1$ (thereby introducing lack of identifiability into the description of ARMA$(1,1)$ autocovariance functions). So it is completely standard to restrict not only to parameters $|\theta|<1$ but also parameters $|\phi|<1$ making the representations of autocovariance functions both "invertible" AND "causal."

2.4 Sample Means, Autocovariances, and Autocorrelations

We next consider what of interest and practical use can be said about natural statistics computed from the realizable/observable vector

$$\boldsymbol{Y}_n = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

(a sub-vector of the conceptually infinite $\boldsymbol{Y}$) under a second order stationary model. We begin with

$$\bar{y}_n = \frac{1}{n}\sum_{t=1}^{n} y_t$$

Clearly

$$\mathrm{E}\bar{y}_n = \mu$$

As it turns out,

$$\begin{aligned} \mathrm{Var}\bar{y}_n &= \frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\mathrm{Cov}\left(y_t, y_s\right) \\ &= \frac{1}{n^2}\sum_{t-s=-n}^{n}\left(n-|t-s|\right)\gamma\left(t-s\right) \\ &= \frac{1}{n}\sum_{s=-n}^{n}\left(1-\frac{|s|}{n}\right)\gamma\left(s\right) \end{aligned}$$

This latter implies that if $\gamma(s)$ converges to 0 fast enough so that $\sum_{s=-n}^{n}|\gamma(s)|$ converges (i.e. $\sum_{s=0}^{\infty}|\gamma(s)| < \infty$), then $\mathrm{Var}\bar{y}_n \to 0$ and thus $\bar{y}_n$ is a "consistent" estimator of μ.

Further, for many second order stationary models (including Gaussian ones, linear and ARMA processes) people can prove central limit results that say that

$$\sqrt{n}\left(\bar{y}_n - \mu\right)$$

is approximately normal for large n. This means that limits

$$\bar{y}_n \pm z\sqrt{\frac{\sum_{s=-n}^{n}\left(1-\frac{|s|}{n}\right)\gamma\left(s\right)}{n}} \tag{22}$$

can (in theory) serve as large sample confidence limits for μ. In applications the sum under the root in display (22) will not be known and will need to be estimated. To this end note that in cases where $\sum_{s=0}^{\infty}|\gamma(s)| < \infty$,

$$\sum_{s=-n}^{n}\left(1-\frac{|s|}{n}\right)\gamma\left(s\right) \to \sum\gamma\left(s\right)$$

So if $\hat{\gamma}_n(s)$ is some estimator of $\gamma(s)$ based on $\mathbf{Y}_n$, a plausible replacement for $\sum_{s=-n}^{n}\left(1-\frac{|s|}{n}\right)\gamma(s)$ in limits (22) is

$$\sum_{s=-\#}^{\#}\hat{\gamma}_n(s)$$

for # chosen so that one can be fairly sure that $\sum_{s=-\#}^{\#}\gamma(s)\approx\sum\gamma(s)$ AND $\hat{\gamma}_n(s)$ is a fairly reliable estimator of $\gamma(s)$ for $|s|\leq\#$. BDM recommends the use of $\#=\sqrt{n}$ and thus realizable approximate limits for μ of the form

$$\bar{y}_n \pm z\sqrt{\frac{\sum_{s=-\sqrt{n}}^{\sqrt{n}}\hat{\gamma}_n(s)}{n}}$$

Regarding sample/estimated covariances, it is standard to define

$$\hat{\gamma}_n(s)\equiv\frac{1}{n}\sum_{t=1}^{n-|s|}(y_t-\bar{y}_n)\left(y_{t+|s|}-\bar{y}_n\right) \tag{23}$$

There are only $n-|s|$ products of the form $(y_t-\bar{y}_n)\left(y_{t+|s|}-\bar{y}_n\right)$ and one might thus expect to see an $n-|s|$ divisor (or some even smaller divisor) in formula (23). But using instead the n divisor is a way of ensuring that a corresponding estimated covariance matrix is non-negative definite. That is, with definition (23) for any $1\leq k\leq n$, the $k\times k$ matrix

$$\hat{\mathbf{\Gamma}}_k=\begin{pmatrix}\hat{\gamma}_n(0) & \hat{\gamma}_n(1) & \cdots & \hat{\gamma}_n(k-1)\\ \hat{\gamma}_n(-1) & \hat{\gamma}_n(0) & \cdots & \hat{\gamma}_n(k-2)\\ \vdots & \vdots & \ddots & \vdots\\ \hat{\gamma}_n(-(k-1)) & \hat{\gamma}_n(-(k-2)) & \cdots & \hat{\gamma}_n(0)\end{pmatrix}$$

is nonnegative definite. In fact, if $\hat{\gamma}_n(0)>0$ (the entries of $\mathbf{Y}_n$ are not all the same) then $\hat{\mathbf{\Gamma}}_k$ is non-singular and therefore positive definite.

The estimated/sample autocorrelation function $\rho(s)$ derived from the values (23) is

$$\hat{\rho}_n(s)\equiv\frac{\hat{\gamma}_n(s)}{\hat{\gamma}_n(0)}$$

Values of this are not very reliable unless n is reasonably large and s is small relative to n. BDM offer the rule of thumb that $\hat{\rho}_n(s)$ should be trusted only when $n\geq 50$ and $|s|\leq n/4$.

Distributional properties of $\hat{\rho}_n(s)$ form the basis for some kinds of inferences for the autocorrelation function. For example, BDM page 61 says that for large n and $\boldsymbol{\rho}_k$ the k-vector $(\rho(1),\rho(2),\ldots,\rho(k))'$, the corresponding vector of sample correlations is approximately multivariate normal, that is

$$\begin{pmatrix}\hat{\rho}_n(1)\\ \hat{\rho}_n(2)\\ \vdots\\ \hat{\rho}_n(k)\end{pmatrix}\dot{\sim}\,\mathrm{MVN}_k\left(\boldsymbol{\rho}_k,\frac{1}{n}\boldsymbol{W}\right)$$

for $\boldsymbol{W}$ a $k \times k$ matrix with (i,j) entry given by "the Bartlett formula"

$$w_{ij} = \sum_{k=1}^{\infty} \{\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\} \{\rho(k+j) + \rho(k-j) - 2\rho(i)\rho(j)\}$$

Note that in particular, the Bartlett formula gives

$$w_{jj} = \sum_{k=1}^{\infty} (\rho(k+j) + \rho(k-j) - 2\rho(i)\rho(j))^2$$

and one can expect $\hat{\rho}_n(s)$ to typically be within, say, $2\sqrt{w_{ss}}/\sqrt{n}$ of $\rho(s)$.

This latter fact can be used as follows. If I have in mind a particular second order stationary model and corresponding autocorrelation function $\rho(s)$, values $\hat{\rho}_n(s)$ (for $|s|$ not too big) outside approximate probability limits

$$\rho(s) \pm 2\frac{\sqrt{w_{ss}}}{\sqrt{n}}$$

suggest lack of fit of the model to a data set in hand. One particularly important application of this is the case of white noise, for which $\rho(0) = 1$ and $\rho(s) = 0$ for $s \neq 0$. It's easy enough to argue that in this case $w_{ss} = 1$ for $s \neq 0$. So a plot of values $\hat{\rho}_n(s)$ versus s with limits drawn on it at

$$\pm 2\frac{1}{\sqrt{n}}$$

is popular as a tool for identifying lags in a time series at which there are detectably non-zero autocorrelations and evidence against the appropriateness of a white noise model.

More generally, if I have in mind a particular pair of orders (p,q) for an ARMA model, I thereby have in mind a functional form for $\rho(s)$ depending upon vector parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, say $\rho_{\boldsymbol{\phi},\boldsymbol{\theta}}(s)$ and therefore values w_{ss} also depending upon $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, say $w_{ss,\boldsymbol{\phi},\boldsymbol{\theta}}$. If I estimate $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ from $\boldsymbol{Y}_n$ as say $\hat{\boldsymbol{\phi}}_n$ and $\hat{\boldsymbol{\theta}}_n$, then I expect $\hat{\rho}_n(s)$ to typically be inside limits

$$\rho_{\hat{\boldsymbol{\phi}}_n,\hat{\boldsymbol{\theta}}_n}(s) \pm 2\frac{\sqrt{w_{ss,\hat{\boldsymbol{\phi}}_n,\hat{\boldsymbol{\theta}}_n}}}{\sqrt{n}} \tag{24}$$

When this fails to happen for small to moderate $|s|$ there is evidence of lack of fit of an ARMA(p,q) model. (This is a version of what is being portrayed on BDM page 63, though for reasons I don't quite understand, the authors draw limits around $\hat{\rho}_n(s)$ and look for $\rho_{\hat{\boldsymbol{\phi}}_n,\hat{\boldsymbol{\theta}}_n}(s)$ values outside those limits rather than vice versa.) Limits (24) are some kind of *very* approximate prediction limits for $\hat{\rho}_n(s)$ ... if one had $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ and used them above, the limits would already be approximate prediction limits because of the reliance upon the large sample normal approximation for the distribution of $\hat{\rho}_n(s)$.

2.5 Prediction and Gaussian Conditional Distributions

Time series models are usually fit for purposes of making predictions of future values of the series. The mathematical formulation of this enterprise is typically "best linear prediction." To introduce this methodology, consider the following (at this point, abstractly stated) problem. For $\underset{k\times 1}{\boldsymbol{V}}$ and $\underset{1\times 1}{u}$ random vectors with

$$\mathrm{E}\begin{pmatrix} \boldsymbol{V} \\ u \end{pmatrix} = \begin{pmatrix} \underset{k\times 1}{\boldsymbol{\mu}_1} \\ \underset{1\times 1}{\mu_2} \end{pmatrix} \quad \text{and} \quad \mathrm{Cov}\begin{pmatrix} \boldsymbol{V} \\ u \end{pmatrix} = \begin{pmatrix} \underset{k\times k}{\boldsymbol{\Sigma}_{11}} & \underset{k\times 1}{\boldsymbol{\Sigma}_{12}} \\ \underset{1\times k}{\boldsymbol{\Sigma}_{21}} & \underset{1\times 1}{\boldsymbol{\Sigma}_{22}} \end{pmatrix}$$

what linear form $c + \boldsymbol{l}'\boldsymbol{V}$ minimizes

$$\mathrm{E}\left(u - \left(c + \boldsymbol{l}'\boldsymbol{V}\right)\right)^2 \tag{25}$$

over choices of $c \in \Re$ and $\boldsymbol{l} \in \Re^k$?

It turns out that this linear prediction question is related to another question whose answer involves basics of probability theory, including conditional means and multivariate normal distributions. That is this. If one adds to the above mean and covariance assumptions the assumption of $(k+1)$-dimensional normality, what function of $\boldsymbol{V}$, say $g(\boldsymbol{V})$, minimizes

$$\mathrm{E}\left(u - g(\boldsymbol{V})\right)^2 \tag{26}$$

over choices of $g(\cdot)$, linear or non-linear? Basic probability facts about conditional distributions and conditional means say that (*in general*) the optimal $g(\boldsymbol{V})$ is

$$\mathrm{E}[u|\boldsymbol{V}]$$

Multivariate normal facts then imply that for Gaussian models

$$\mathrm{E}[u|\boldsymbol{V}] = \mu_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{V} - \boldsymbol{\mu}_1) \tag{27}$$

Now this normal conditional mean (27) is in fact a linear form $c + \boldsymbol{l}'\boldsymbol{V}$ (with $c = \mu_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1$ and $\boldsymbol{l}' = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$). So since it optimizes the more general criterion (26) it also optimizes the original criterion (25) for normal models. But the original criterion takes the same value for all second order stationary models with a given moment structure, regardless of whether or not a model is Gaussian. That means that the form (27) is the solution to the original linear prediction problem in general.

Note also that for the multivariate normal model,

$$\mathrm{Var}[u|\boldsymbol{V}] = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \tag{28}$$

(an expression that we note does not depend upon the value of $\boldsymbol{V}$). What is also interesting about the form of the normal conditional variance (28) is that

it gives the optimal value of prediction mean square error (26). To see this, note that in general

$$\begin{aligned}
\mathrm{E}\left(u - \mathrm{E}\left[u|\boldsymbol{V}\right]\right)^2 &= \mathrm{E}\left(\left(u - \mathrm{E}u\right) - \left(\mathrm{E}\left[u|\boldsymbol{V}\right] - \mathrm{E}u\right)\right)^2 \\
&= \mathrm{E}\left(u - \mathrm{E}u\right)^2 + \mathrm{Var}\ \mathrm{E}\left[u|\boldsymbol{V}\right] - 2\mathrm{E}\left(\left(u - \mathrm{E}u\right)\left(\mathrm{E}\left[u|\boldsymbol{V}\right] - \mathrm{E}u\right)\right) \\
&= \mathrm{Var}u + \mathrm{Var}\ \mathrm{E}\left[u|\boldsymbol{V}\right] - 2\mathrm{E}\left(\mathrm{E}\left[\left(u - \mathrm{E}u\right)\left(\mathrm{E}\left[u|\boldsymbol{V}\right] - \mathrm{E}u\right)|\boldsymbol{V}\right]\right) \\
&= \mathrm{Var}u - \mathrm{Var}\ \mathrm{E}\left[u|\boldsymbol{V}\right]
\end{aligned}$$

Then in normal cases, $\mathrm{E}(u - \mathrm{E}\left[u|\boldsymbol{V}\right])^2$ on the left above is the minimum value of criterion (25), and since $\mathrm{E}[u|\boldsymbol{V}]$ has form (27) basic probability facts for linear combinations of random variables imply that

$$\begin{aligned}
\mathrm{Var}\ \mathrm{E}\left[u|\boldsymbol{V}\right] &= \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathrm{Cov}\left(\boldsymbol{V} - \boldsymbol{\mu}_1\right)\left(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\right)' \\
&= \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\
&= \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}
\end{aligned}$$

So indeed, the minimum value of criterion (25) in normal cases is

$$\mathrm{Var}u - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \tag{29}$$

the normal conditional variance (28). Since this is true for normal cases and the value of criterion (25) is the same for every model with the given mean and covariance structure, $\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ is then the minimum value of criterion (25) for any model with this mean and covariance structure.

All of the above generality can be applied to various prediction questions for weakly stationary time series, with $\boldsymbol{V}$ some finite part of $\boldsymbol{Y}$ and u some coordinate of $\boldsymbol{Y}$ not in $\boldsymbol{V}$. (We'll actually have reason in a bit to consider u of dimension larger than 1, but for the time being stick with scalar u.)

Consider first the prediction of y_{n+s} based on $\boldsymbol{Y}_n$. The vector

$$\begin{pmatrix} \boldsymbol{Y}_n \\ y_{n+s} \end{pmatrix}$$

has

$$\mathrm{E}\begin{pmatrix} \boldsymbol{Y}_n \\ y_{n+s} \end{pmatrix} = \mu_{\underset{(n+1)\times 1}{\mathbf{1}}}$$

and

$$\mathrm{Cov}\begin{pmatrix} \boldsymbol{Y}_n \\ y_{n+s} \end{pmatrix} = \begin{pmatrix}
\gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(n-1) & \gamma(n+s-1) \\
\gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(n-2) & \gamma(n+s-2) \\
\gamma(2) & \gamma(1) & \gamma(0) & \cdots & \gamma(n-3) & \gamma(n+s-3) \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\gamma(n-1) & \gamma(n-2) & \gamma(n-3) & \cdots & \gamma(0) & \gamma(s) \\
\gamma(n+s-1) & \gamma(n+s-2) & \gamma(n+s-3) & \cdots & \gamma(s) & \gamma(0)
\end{pmatrix}$$

from whence we may define

$$\underset{n\times n}{\boldsymbol{\Sigma}_{11}} = \begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(n-2) \\ \gamma(2) & \gamma(1) & \gamma(0) & \cdots & \gamma(n-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \gamma(n-3) & \cdots & \gamma(0) \end{pmatrix},$$

$$\underset{n\times 1}{\boldsymbol{\Sigma}_{12}} = \begin{pmatrix} \gamma(n+s-1) \\ \gamma(n+s-2) \\ \gamma(n+s-3) \\ \vdots \\ \gamma(s) \end{pmatrix}, \boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}' \text{ and } \boldsymbol{\Sigma}_{22} = \gamma(0)$$

Then (using BDM notation) with

$$P_n y_{n+s} = \text{the best linear predictor of } y_{n+s} \text{ from } \boldsymbol{Y}_n$$

it is the case that

$$P_n y_{n+s} = \mu + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{Y}_n - \mu\mathbf{1}) \tag{30}$$

In some sense, application of this development to various specific second order stationary models (each with a different autocovariance function $\gamma(s)$) is then "just" a matter of details. $P_n y_{n+s}$ may have special nice forms for some models, and others may present really nasty computational problems in order to actually compute $P_n y_{n+s}$, but that is all in the realm of the specialist. For users, what is important is the big picture that says this is all just use of a multivariate normal form.

At this point it is probably important to stop and say that in applications, best linear prediction (depending as it does on the mean and covariance structure that can only be learned from data) is not realizable. That is, in order to use form (30) one must know μ and $\gamma(s)$. In practice, the best one will be able to muster are estimates of these. But if, for example, one fits an ARMA(p,q) model, producing estimates of (possibly a non-zero mean μ and) parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$ and σ, these can be plugged into an ARMA(p,q) form for $\gamma(s)$ to produce estimated $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$ and then an approximate $P_n y_{n+s}$, say $\widehat{P_n y_{n+s}}$. (This, by the way, is very parallel to the story about "BLUPs" told in a course like Stat 511. One cannot actually use the optimal c and $\boldsymbol{l}'$, but can at least hope to estimate them without too much loss, and produce good "approximate BLUPs" $\widehat{P_n y_{n+s}}$.)

Note also that IF one did have available the actual autocovariance function, under multivariate normality, the limits

$$P_n y_{n+s} \pm z\sqrt{\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}}$$

would function as (theoretically exact) prediction limits for y_{n+s}. Having to estimate model parameters of some autocovariance function (and perhaps a

mean) to produce realizable limits

$$\widehat{P_n y_{n+s}} \pm z\sqrt{\left(\boldsymbol{\Sigma}_{22} - \widehat{\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}}\right)}$$

makes these surely approximate and potentially substantially optimistic, since this form fails to take account of the "extra" uncertainty in the prediction associated with the fact that the parameter estimates are imperfect/noisy.

BDM Section 2.5 has a number of results concerning "the prediction operator" $P_n(\cdot)$ that might be used to prove things about prediction and find tricks that can simplify computations in special models. It seems to me that instead of concerning oneself with those results in and of themselves, it makes more sense to simply make use of the "conditional mean operator" $\mathrm{E}[\cdot|\boldsymbol{Y}_n]$ for a Gaussian version of a second order stationary model, and then note that whatever is true involving it is equally true concerning $P_n(\cdot)$ in general. Some applications of this way of operating follow.

Consider, for example, an AR(1) model with mean 0 and prediction in that model. That is, consider a models specified by

$$y_t = \phi y_{t-1} + \epsilon_t$$

for $\boldsymbol{\epsilon}$ white noise and $|\phi| < 1$. If the ϵ_t are jointly Gaussian, ϵ_t is independent of $(\ldots, \epsilon_{t-3}, \epsilon_{t-2}, \epsilon_{t-1})$ and y_{t-1} is a function of this infinite set of variables. Consider then

$$P_n y_{n+1}$$

the one-step-ahead forecast. For a Gaussian model

$$\begin{aligned} P_n y_{n+1} &= \mathrm{E}[y_{n+1}|\boldsymbol{Y}_n] \\ &= \mathrm{E}[\phi y_n + \epsilon_{n+1}|\boldsymbol{Y}_n] \\ &= \phi\mathrm{E}[y_n|\boldsymbol{Y}_n] + \mathrm{E}[\epsilon_{n+1}|\boldsymbol{Y}_n] \\ &= \phi y_n + 0 \\ &= \phi y_n \end{aligned}$$

(because of the linearity of conditional expectation, the fact that y_n is a function of $\boldsymbol{Y}_n$, and ϵ_{n+1} has mean 0 and is independent of $(\ldots, \epsilon_{n-3}, \epsilon_{n-2}, \epsilon_{n-1}, \epsilon_n)$ and therefore $\boldsymbol{Y}_n$). $P_n y_{n+1} = \phi y_n$ being the case for Gaussian AR(1) models means it's true for all AR(1) models.

More generally, it is the case that for AR(1) models with mean 0 and $|\phi| < 1$,

$$P_n y_{n+s} = \phi^s y_n \tag{31}$$

and the corresponding prediction variance (29) is the Gaussian $\mathrm{Var}[y_{n+s}|\boldsymbol{Y}_n]$

$$\sigma^2 \frac{1-\phi^{2s}}{1-\phi^2}$$

That the AR(1) s-step-ahead predictor is as in display (31) follows as for the $s = 1$ case after using the recursion to write

$$y_{n+s} = \phi^s y_n + \phi^{s-1}\epsilon_{n+1} + \phi^{s-2}\epsilon_{n+2} + \cdots + \phi\epsilon_{n+s-1} + \epsilon_{n+s}$$

It is also worth considering the relationship of prediction for a mean 0 process to that of a mean μ (possibly non-zero) process. If $\boldsymbol{Y}$ is second order stationary with mean vector $\mu\mathbf{1}$, then

$$\boldsymbol{Y}^* = \boldsymbol{Y} - \mu\mathbf{1}$$

is a mean 0 process with the same autocovariance function as $\boldsymbol{Y}$. Then for a Gaussian version of these models

$$\begin{aligned} P_n y_{n+s} &= P_n\left(y^*_{n+s} + \mu\right) \\ &= \mathrm{E}\left[y^*_{n+s} + \mu | \boldsymbol{Y}_n\right] \\ &= \mathrm{E}\left[y^*_{n+s} | \boldsymbol{Y}_n\right] + \mu \\ &= \mathrm{E}\left[y^*_{n+s} | \boldsymbol{Y}_n - \mu\mathbf{1}\right] + \mu \\ &= \mathrm{E}\left[y^*_{n+s} | \boldsymbol{Y}^*_n\right] + \mu \\ &= \mu + P^*_n y^*_{n+s} \end{aligned}$$

(the fourth equality following because conditioning on the value of $\boldsymbol{Y}_n$ is no different from conditioning on the value of $\boldsymbol{Y}_n - \mu\mathbf{1}$, knowing the value of one is exactly equivalent to knowing the value of the other). The notation $P^*_n y^*_{n+s}$ is meant to indicate the prediction operator for the mean 0 version of the process based on $\boldsymbol{Y}^*_n$ applied to y^*_{n+s}. Since the first element in the string of equalities is the same as the last for Gaussian processes, it is the same for all second order processes. So knowing how to predict a mean 0 process, one operates on values for the original series with μ subtracted to predict a (μ-subtracted) future value and then adds μ back in. For example, for an AR(1) process with mean μ, the s-step-ahead forecast for y_{n+s} based on $\boldsymbol{Y}_n$ is

$$\mu + \phi^s y^*_n = \mu + \phi^s\left(y_n - \mu\right)$$

In light of the simple relationship between forecasts for mean 0 processes and for mean μ processes (and the fact that much of time series analysis is about forecasting) it is standard to assume that the mean is 0 unless explicitly stated to the contrary, and we'll adopt that convention for the time being.

2.6 Partial Autocorrelations

One additional practically important concept naturally related to the use of Gaussian assumptions to generate general formulas for second order stationary time series models is that of **partial autocorrelation**. It derives most naturally from the multivariate normal conditioning formulas, not for a univariate quantity, but for a bivariate quantity. Consider a Gaussian stationary process

and the finite vector

$$\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_s \end{pmatrix}$$

The multivariate normal conditioning formulas tell how to describe the conditional distribution

$$\begin{pmatrix} y_0 \\ y_s \end{pmatrix} \text{ given } \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{s-1} \end{pmatrix}$$

Rearranging slightly for convenience, the vector

$$\begin{pmatrix} y_1 \\ \vdots \\ y_{s-1} \\ y_s \\ y_0 \end{pmatrix}$$

has mean $\mathbf{0}$ and covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(s-3) & \gamma(s-2) & \gamma(s-1) & \gamma(1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(s-4) & \gamma(s-3) & \gamma(s-2) & \gamma(2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \gamma(s-3) & \gamma(s-4) & \cdots & \gamma(0) & \gamma(1) & \gamma(2) & \gamma(s-2) \\ \gamma(s-2) & \gamma(s-3) & \cdots & \gamma(1) & \gamma(0) & \gamma(1) & \gamma(s-1) \\ \gamma(s-1) & \gamma(s-2) & \cdots & \gamma(2) & \gamma(1) & \gamma(0) & \gamma(s) \\ \gamma(1) & \gamma(2) & \cdots & \gamma(s-2) & \gamma(s-1) & \gamma(s) & \gamma(0) \end{pmatrix}$$

which we partition as

$$\mathbf{\Sigma} = \begin{pmatrix} \underset{(s-1)\times(s-1)}{\mathbf{\Sigma}_{11}} & \underset{(s-1)\times 2}{\mathbf{\Sigma}_{12}} \\ \underset{2\times(s-1)}{\mathbf{\Sigma}_{21}} & \underset{2\times 2}{\mathbf{\Sigma}_{22}} \end{pmatrix}$$

for

$$\mathbf{\Sigma}_{11} = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(s-3) & \gamma(s-2) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(s-4) & \gamma(s-3) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(s-3) & \gamma(s-4) & \cdots & \gamma(0) & \gamma(1) \\ \gamma(s-2) & \gamma(s-3) & \cdots & \gamma(1) & \gamma(0) \end{pmatrix}, \mathbf{\Sigma}_{22} = \begin{pmatrix} \gamma(0) & \gamma(s) \\ \gamma(s) & \gamma(0) \end{pmatrix},$$

$$\mathbf{\Sigma}_{21} = \begin{pmatrix} \gamma(s-1) & \gamma(s-2) & \cdots & \gamma(2) & \gamma(1) \\ \gamma(1) & \gamma(2) & \cdots & \gamma(s-2) & \gamma(s-1) \end{pmatrix}, \text{ and } \mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21}'$$

Then, the conditional covariance matrix for $(y_0, y_s)'$ is (in the same basic form as used repeatedly above, but now a 2×2 matrix)

$$\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

This prescribes a conditional covariance between y_0 and y_s given the intervening observations, and then a conditional correlation between them. We'll use the notation (following BDM)

$$\begin{aligned} \alpha(s) &= \text{the Gaussian conditional correlation between} \\ &\quad\ y_0 \text{ and } y_s \text{ given } y_1, y_2, \ldots, y_{s-1} \end{aligned}$$

and call $\alpha(s)$ the **partial autocorrelation function** for the model. Notice that by virtue of the fact that normal conditional covariance matrices do not depend upon the values of conditioning variables, $\alpha(s)$ is truly a function only of the lag, s, involved and the form of the autocovariance function $\gamma(s)$.

It remains to provide some motivation/meaning for $\alpha(s)$ outside of the Gaussian context. BDM provide several kinds of help in that direction, one computational and others that are more conceptual. In the first place, they point out that for any second order stationary process, with

$$\boldsymbol{\Gamma}_s = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(s-1) \\ \gamma(-1) & \gamma(0) & \cdots & \gamma(s-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(-(s-1)) & \gamma(-(s-2)) & \cdots & \gamma(0) \end{pmatrix} \quad \text{and} \quad \boldsymbol{\gamma}_s = \begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(s) \end{pmatrix}$$

$$\alpha(s) = \text{the } s\text{th entry of } \boldsymbol{\Gamma}_s^{-1}\boldsymbol{\gamma}_s \tag{32}$$

Further, it is the case that in general

1. $\alpha(s)$ is the correlation between the (linear) prediction errors $y_s - P_{s-1}y_s$ and $y_0 - P_{s-1}y_0$ (for P_{s-1} the linear prediction operator based on $\boldsymbol{Y}_{s-1}$),
2. for $P_s y_0 = c + \sum_{t=1}^{s} l_t y_t$ the best linear predictor of y_0 based on $\boldsymbol{Y}_s$, $\alpha(s) = l_s$, and
3. for v_n the optimal 1-step-ahead prediction variance based on $\boldsymbol{Y}_n$, $v_n = \mathrm{E}(y_{n+1} - P_n y_{n+1})^2$ it is the case that

   $$v_n = v_{n-1}\left(1 - \alpha(n)^2\right)$$

   (so that the larger is $|\alpha(n)|$ the greater is the reduction in prediction variance associated with an increase of 1 in the length of the data record available for use in prediction).

A primary use of an estimated partial autocorrelation function (derived, for example from estimated versions of relationship (32)) is in model identification. For example, an AR(p) model has $\alpha(s) = 0$ for $s > p$. So a sample partial autocorrelation function that is very small at lags larger than p suggests the possibility that an AR(p) might fit a data set in hand.

3 General ARMA(p,q) Models

3.1 ARMA Models and Some of Their Properties

It's fairly clear how one proceeds to generalize the AR(1),MA(q), and ARMA(1, 1) models of the previous section. For backshift polynomial operators

$$\Phi(\mathcal{B}) = \mathcal{I} - \phi_1\mathcal{B} - \phi_2\mathcal{B}^2 - \cdots - \phi_p\mathcal{B}^p$$

and

$$\Theta(\mathcal{B}) = \mathcal{I} + \theta_1\mathcal{B} + \theta_2\mathcal{B}^2 + \cdots + \theta_q\mathcal{B}^q$$

and white noise process $\boldsymbol{\epsilon}$, we consider the possibility of a time series $\boldsymbol{Y}$ satisfying the ARMA(p,q) equation

$$\Phi(\mathcal{B})\,\boldsymbol{Y} = \Theta(\mathcal{B})\,\boldsymbol{\epsilon} \tag{33}$$

Of course, in notation not involving operators, this is

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}\ \forall t \tag{34}$$

In order to represent a $\boldsymbol{Y}$ solving equation (33) as a causal time-invariant linear process, one wants the operator $\Phi(\mathcal{B})$ to be invertible. As it turns out, a standard argument (provided most clearly on page 85 of BDT) says that $\Phi(\mathcal{B})$ has an inverse provided the polynomial

$$\phi(z) \equiv 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$$

treated as a map from the complex numbers to the complex numbers has no roots inside the unit circle (i.e. if $|z| < 1$ then $\phi(z) \neq 0$). In that event, there is a causal time invariant linear operator $\mathcal{L}$ for which

$$\boldsymbol{Y} = \mathcal{L}\Theta(\mathcal{B})\,\boldsymbol{\epsilon}$$

and it turns out that provided the polynomial

$$\theta(z) \equiv 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_p z^p$$

and $\phi(z)$ have no roots in common, the coefficients ψ_s of

$$\mathcal{L}\Theta(\mathcal{B}) = \sum_{s=0}^{\infty}\psi_s\mathcal{B}^s$$

are such that

$$\sum_{s=0}^{\infty}\psi_s z^s = \frac{\theta(z)}{\phi(z)}\quad \forall\,|z| < 1$$

(This is no computational prescription for the coefficients, but does suggest that they are probably computable.)

It should further be plausible that to the extent that "invertibility" (the ability to write $\boldsymbol{\epsilon}$ in terms of a causal linear filter applied to $\boldsymbol{Y}$) of the process

is of interest, one wants the operator $\Theta(\mathcal{B})$ to have an inverse. Applying the same technical development that guarantees the invertibility of $\Phi(\mathcal{B})$ (see page 87 of BDT) one has that $\Theta(\mathcal{B})$ has an inverse provided the polynomial $\theta(z)$ has no roots inside the unit circle (i.e. if $|z| < 1$ then $\theta(z) \neq 0$). In that event, there is a causal time-invariant linear operator $\mathcal{M}$ for which

$$\boldsymbol{\epsilon} = \mathcal{M}\Phi(\mathcal{B})\,\boldsymbol{Y}$$

and provided the polynomial $\theta(z)$and $\phi(z)$ have no roots in common, the coefficients π_s of

$$\mathcal{M}\Phi(\mathcal{B}) = \sum_{s=0}^{\infty} \pi_s \mathcal{B}^s$$

are such that

$$\sum_{s=0}^{\infty} \pi_s z^s = \frac{\phi(z)}{\theta(z)} \quad \forall\, |z| < 1$$

Given the above development, it is not surprising that in order to avoid identifiability problems when estimating the parameters of ARMA models people commonly restrict attention to coefficient sets for which neither $\phi(z)$ nor $\theta(z)$ have roots inside the unit circle and the polynomials have no common factors, so the corresponding stationary solutions to the ARMA(p,q) equation (33) are both causal and invertible.

Computation of the coefficients ψ_s for $\mathcal{L}\Theta(\mathcal{B}) = \sum_{s=0}^{\infty} \psi_s \mathcal{B}^s$ can proceed recursively by equating coefficients in the power series identity

$$\phi(z) \sum_{s=0}^{\infty} \psi_s z^s = \theta(z)$$

i.e.

$$\left(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p\right)\left(\psi_0 + \psi_1 z^1 + \psi_2 z^2 + \cdots\right) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_p z^p$$

That is, clearly

$$\begin{aligned} \psi_0 &= 1 \\ \psi_1 &= \theta_1 + \psi_0\phi_1 \\ \psi_2 &= \theta_2 + \psi_1\phi_1 + \psi_0\phi_2 \end{aligned}$$

and in general

$$\begin{aligned} \psi_j &= \theta_j + \sum_{k=1}^{p} \phi_k \psi_{j-k} \quad \text{for } j = 0, 1, \ldots \qquad (35) \\ \text{where } \theta_0 &= 1, \theta_j = 0 \quad \text{for } j > q \quad \text{and } \psi_j = 0 \quad \text{for } j < 0 \end{aligned}$$

Of course, where one has use for the impulse response function for $\mathcal{M}\Phi(\mathcal{B}) = \sum_{s=0}^{\infty} \pi_s \mathcal{B}^s$, similar reasoning produces a recursion

$$\begin{aligned} \pi_j &= -\phi_j - \sum_{k=1}^{q} \theta_k \pi_{j-k} \quad \text{for } j = 0, 1, \ldots \\ \text{where } \phi_0 &= -1, \phi_j = 0 \quad \text{for } j > p \quad \text{and } \pi_j = 0 \quad \text{for } j < 0 \end{aligned} \tag{36}$$

One reason for possibly wanting the weights π_j in practice is this. An ARMA process satisfying not the equation (33), but rather

$$\Theta(\mathcal{B})\, \boldsymbol{Y} = \Phi(\mathcal{B})\, \boldsymbol{\epsilon}$$

is obviously related to the original ARMA(p, q) in some way. As it turns out, a common model identification tool is the autocorrelation function of this "dual process," sometimes called the **inverse autocorrelation function**. Since the MA(q) autocorrelations for lags larger than q are 0, it follows that an AR(p) process has an inverse autocorrelation function that is 0 beyond lag p. So looking at an inverse autocorrelation function is an alternative to considering the partial autocorrelation function to identify an AR process and its order.

3.2 Computing ARMA(p, q) Autocovariance Functions and (Best Linear) Predictors

Once one has coefficients ψ_t for representing $\boldsymbol{Y} = \mathcal{L}\Theta(\mathcal{B})\, \boldsymbol{\epsilon}$ as a linear process, expression (10) immediately provides a form for the autocovariance function, namely

$$\gamma(s) = \sigma^2 \sum_{t=-\infty}^{\infty} \psi_t \psi_{t+s}$$

This form is not completely happy, involving as it does an infinite series and all the coefficients ψ_t. But there is another insight that allows efficient computation of the autocovariance function.

If one multiplies the basic ARMA equation (34) through by y_{t-k} and takes expectations, the relationship

$$\begin{aligned} & \gamma(k) - \phi_1 \gamma(k-1) - \phi_2 \gamma(k-2) - \cdots - \phi_p \gamma(k-p) \\ = \quad & \mathrm{E}\left[\left(\sum_{s=0}^{\infty} \psi_s \epsilon_{t-k-s} \right) \left(\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \right) \right] \end{aligned}$$

is produced. The right hand side of this equation is (since $\boldsymbol{\epsilon}$ is white noise)

$$\begin{array}{ll} \sigma^2 \left(\psi_0 + \theta_1 \psi_1 + \cdots + \theta_q \psi_q \right) & \text{for } k = 0 \\ \sigma^2 \left(\theta_1 \psi_0 + \theta_2 \psi_1 + \cdots + \theta_q \psi_{q-1} \right) & \text{for } k = 1 \\ \sigma^2 \left(\theta_2 \psi_0 + \theta_3 \psi_1 + \cdots + \theta_q \psi_{q-2} \right) & \text{for } k = 2 \\ \vdots & \vdots \end{array}$$

This suggests that the first $p+1$ of these equations (the ones for $k = 0, 1, \ldots, p$) may be solved simultaneously for $\gamma(0), \gamma(1), \ldots, \gamma(p)$ and then one may solve recursively for

$$\begin{array}{ll} \gamma(p+1) & \text{using the } k = p+1 \text{ equation} \\ \gamma(p+2) & \text{using the } k = p+2 \text{ equation} \\ \vdots & \vdots \end{array}$$

This method of computing, doesn't require approximating an infinite sum and requires computation of only $\psi_0, \psi_1, \ldots, \psi_q$ from the parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ^2.

In theory, prediction for the ARMA(p, q) process can follow the basic "mean of the conditional normal distribution" path laid down in Section 2.5. But as a practical matter, direct use of that development requires inversion of the $n \times n$ matrix $\boldsymbol{\Sigma}_{11}$ in order to compute predictions, and that would seem to get out of hand for large n. A way out of this matter is through the use of the so-called **innovations** (one-step-ahead prediction errors) **algorithm**. This is first discussed in general terms in Section 2.5.2 of BDM. Before here considering its specialization to ARMA models, we consider a theoretical development that points in the direction of the algorithm.

Temporarily consider a causal invertible Gaussian ARMA(p, q) model and write

$$y_t = \sum_{s=0}^{\infty} \psi_s \epsilon_{t-s} \quad \text{and} \quad \epsilon_t = \sum_{s=0}^{\infty} \pi_s y_{t-s}$$

so that in theory, knowing the infinite sequence of observations through time n (namely $\ldots y_{-1}, y_0, \ldots, y_n$) is equivalent to knowing the infinite sequence of errors through time n (namely $\ldots \epsilon_{-1}, \epsilon_0, \ldots, \epsilon_n$). In light of the representation (34), a theoretical (non-realizable) predictor of y_{n+1} is

$$\begin{aligned} \mathrm{E}[y_{n+1} | \ldots y_{-1}, y_0, \ldots, y_n] &= \phi_1 y_n + \phi_2 y_{n-1} + \cdots + \phi_p y_{n+1-p} \\ &\quad + 0 + \theta_1 \epsilon_n + \theta_2 \epsilon_{n-1} + \cdots + \theta_q \epsilon_{n+1-q} \end{aligned}$$

(this is not realizable because one doesn't ever have an infinite record $\ldots y_{-1}, y_0, \ldots, y_n$ available and can't recover the ϵ's). In deriving this theoretical predictor, one uses the equivalence of the y and ϵ informations and the fact that for the Gaussian case, ϵ_{n+1} is independent of the conditioning sequence and has mean 0. It is plausible that $P_n y_{n+1}$ might have a form somewhat like this theoretical one and the innovations algorithm shows this is the case.

With

$$\begin{aligned} \hat{y}_1 &\equiv 0 \\ \hat{y}_n &= P_{n-1} y_n \quad \text{for } n = 2, 3, \ldots \\ v_n &= \mathrm{E}(y_{n+1} - \hat{y}_{n+1})^2 \quad \text{for } n = 0, 1, 2, \ldots \end{aligned}$$

the ARMA specialization of the innovations algorithm shows that

$$\hat{y}_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{nj} \left(y_{n+1-j} - \hat{y}_{n+1-j} \right) & 1 \leq n < \max(p,q) \\ \phi_1 y_n + \phi_2 y_{n-1} + \cdots + \phi_p y_{n+1-p} & \\ \quad + \sum_{j=1}^{q} \theta_{nj} \left(y_{n+1-j} - \hat{y}_{n+1-j} \right) & n \geq \max(p,q) \end{cases}$$

and

$$v_n = \sigma^2 r_n$$

where the θ_{nj}'s and the r_n's can be computed recursively using the parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ^2 (and autocovariance function $\gamma(s)$ derived from them, but not involving the observed $\boldsymbol{Y}_n$) and these equations allow recursive computation of $\hat{y}_2, \hat{y}_3, \ldots$. (See BDM pages 100-102 for details.) Note too that if $n > \max(p,q)$ the one-step-ahead forecasts from $\boldsymbol{Y}_n$, have a form much like the theoretical predictor $\mathrm{E}[y_{n+1} | \ldots y_{-1}, y_0, \ldots, y_n]$, where θ_{nj}'s replace θ_j's and innovations replace ϵ's.

Further, once one-step-ahead predictions are computed, they can be used to produce s-step-ahead predictions. That is, in the ARMA(p,q) model

$$P_n y_{n+s} = \begin{cases} \sum_{j=s}^{n+s-1} \theta_{n+s-1,j} \left(y_{n+s-j} - \hat{y}_{n+s-j} \right) & 1 \leq s \leq \max(p,q) - n \\ \sum_{j=1}^{p} \phi_j P_n y_{n+s-j} & \\ \quad + \sum_{j=s}^{n+s-1} \theta_{n+s-1,j} \left(y_{n+s-j} - \hat{y}_{n+s-j} \right) & s > \max(p,q) - n \end{cases}$$

and once $\hat{y}_2, \hat{y}_3, \ldots, \hat{y}_n$ have been computed, for fixed n it's possible to compute $P_n y_{n+1}, P_n y_{n+2}, \ldots$. BDM pages 105-106 also provide prediction variances

$$\begin{aligned} v_n(s) &\equiv \mathrm{E}\left(y_{n+s} - P_n y_{n+s} \right)^2 \\ &= \sum_{j=0}^{s-1} \left(\sum_{r=0}^{j} u_r \theta_{n+s-r-1,j-r} \right)^2 v_{n+s-j-1} \end{aligned}$$

where the coefficients u_j can be computed recursively from

$$u_j = \sum_{k=1}^{\min(p,j)} \phi_k u_{j-k} \quad j = 1, 2, \ldots$$

and (under Gaussian assumptions) prediction limits for y_{n+s} are

$$P_n y_{n+s} \pm z \sqrt{v_n(s)}$$

Of course, where parameters estimates $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}$, and $\widehat{\sigma^2}$ replace $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ^2, the limits

$$\widehat{P_n y_{n+s}} \pm z \sqrt{\widehat{v_n(s)}}$$

are then approximate prediction limits.

3.3 Fitting ARMA(p,q) Models (Estimating Model Parameters)

We now consider basic estimation of the ARMA(p,q) parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ^2 from $\boldsymbol{Y}_n$. The most natural place to start looking for a method of estimation is with maximum likelihood based on a Gaussian version of the model. That is, for $\gamma_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}(s)$ the autocovariance function corresponding to parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ^2 and the $n \times n$ covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2} = \left(\gamma_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}(|i-j|)\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

the Gaussian density for $\boldsymbol{Y}_n$ has the form

$$f\left(\boldsymbol{y}_n|\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2\right) = (2\pi)^{-n/2}\left|\det \boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}\right|^{-n/2} \exp\left(-\frac{1}{2}\boldsymbol{y}_n' \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2} \boldsymbol{y}_n\right)$$

Maximizers $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}$, and $\widehat{\sigma^2}$ of $f\left(\boldsymbol{y}_n|\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2\right)$ are **maximum likelihood** estimates of the parameters. Standard statistical theory then implies that for $H\left(\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2\right)$ the $(p+q+1)\times(p+q+1)$ Hessian matrix (the matrix of second partials) for $\ln f\left(\boldsymbol{y}_n|\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2\right)$,

$$-H^{-1}\left(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}},\widehat{\sigma^2}\right)$$

functions as an estimated covariance matrix for the maximum likelihood estimators, and an estimate $\hat{\phi}$ or $\hat{\theta}$ or $\widehat{\sigma^2}$ plus or minus z times the root of the corresponding diagonal entry of $-H^{-1}\left(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}},\widehat{\sigma^2}\right)$ provides approximate confidence limits for the corresponding parameter.

Direct use of the program just outlined would seem to be limited by the necessity of inverting the $n\times n$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}$ in order to compute the likelihood. What is then helpful is the alternative representation

$$f\left(\boldsymbol{y}_n|\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2\right) = \frac{1}{\sqrt{(2\pi)^n \prod_{j=0}^{n-1} v_j}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(y_j-\hat{y}_j\right)^2/v_{j-1}\right)$$

where the one-step-ahead forecasts $\hat{y}_j$ and prediction variances v_j are functions of the parameters and can be computed as indicated in the previous section. This form enables the proof that with

$$S(\boldsymbol{\phi},\boldsymbol{\theta}) = \sum_{j=1}^{n}\left(y_j-\hat{y}_j\right)^2/r_{j-1}$$

it is the case that

$$\widehat{\sigma^2} = \frac{1}{n}S\left(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}}\right)$$

where $\left(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}\right)$ optimizes

$$l\left(\boldsymbol{\phi}, \boldsymbol{\theta}\right) = \ln\left(\frac{1}{n} S\left(\boldsymbol{\phi}, \boldsymbol{\theta}\right)\right) + \frac{1}{n}\sum_{j=1}^{n} \ln r_{j-1}$$

Further, for $H_1\left(\boldsymbol{\phi}, \boldsymbol{\theta}\right)$ the Hessian matrix for $l\left(\boldsymbol{\phi}, \boldsymbol{\theta}\right)$, an estimated covariance matrix for $\left(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}\right)$ is

$$2H_1^{-1}\left(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}\right)$$

A standard alternative to use of the Gaussian likelihood is **least squares** estimation. In the present situation this is minimization of

$$S\left(\boldsymbol{\phi}, \boldsymbol{\theta}\right) \quad \text{or} \quad \tilde{S}\left(\boldsymbol{\phi}, \boldsymbol{\theta}\right) \equiv \sum_{j=1}^{n} \left(y_j - \hat{y}_j\right)^2$$

to produce estimates $\left(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}\right)$ and then use of the estimate

$$\widetilde{\sigma^2} = \frac{1}{n-p-q} S\left(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}\right) \quad \text{or} \quad \frac{1}{n}\tilde{S}\left(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}\right)$$

The first of these is suggested in BDM and might be termed a kind of "weighted least squares" and the second in Madsen and might be termed "ordinary least squares." With $H_2\left(\boldsymbol{\phi}, \boldsymbol{\theta}\right)$ the Hessian of $\tilde{S}\left(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}\right)$, Madsen says that an estimated covariance matrix for $\left(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}\right)$ is

$$2\tilde{\sigma} H_2^{-1}\left(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}\right)$$

that provides standard errors and then approximate confidence limits for elements of $\left(\boldsymbol{\phi}, \boldsymbol{\theta}\right)$.

A computationally simpler variant of least squares is the **conditional least squares** of Abraham and Ledolter. This is based on the relationship

$$\begin{aligned} \epsilon_t \quad = \quad & y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} \\ & -\theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q} \quad \forall t \end{aligned}$$

obtained by rearranging the basic ARMA relationship (34) (and no doubt motivated by the form of the theoretical predictor $\mathrm{E}[y_{n+1} | \ldots y_{-1}, y_0, \ldots, y_n]$). IF one knew some consecutive string of q values of ϵ's, by observing y's one would know all subsequent ϵ's as well. Thinking that ϵ's have mean 0 and in any case, many periods after a "start-up" string of q values ϵ, the exact values in the start-up string are probably largely immaterial, one might set

$$\widetilde{\epsilon_p} = \widetilde{\epsilon_{p-1}} = \cdots = \widetilde{\epsilon_{p-q+1}} = 0$$

and then compute subsequent approximate ϵ values using

$$\begin{aligned}\widetilde{\epsilon_t} &= y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} \\ &\quad -\theta_1\widetilde{\epsilon_{t-1}} - \theta_2\widetilde{\epsilon_{t-2}} - \cdots - \theta_q\widetilde{\epsilon_{t-q}} \quad \forall t > p\end{aligned}$$

The "conditional least squares" criterion is then

$$S_{\mathrm{C}}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=p+1}^{n} \widetilde{\epsilon_t}^2$$

and minimizers of this criterion $\left(\widetilde{\boldsymbol{\phi}}_{\mathrm{C}}, \widetilde{\boldsymbol{\theta}}_{\mathrm{C}}\right)$ are conditional least squares estimates. For what it is worth, this seems to be the default ARMA fitting method in SAS FS$^{\mathrm{TM}}$. I suspect (but am not actually sure) that for

$$\widetilde{\sigma_{\mathrm{C}}^2} = \frac{1}{n-p} S_{\mathrm{C}}(\boldsymbol{\phi}, \boldsymbol{\theta})$$

and $H_{\mathrm{C}}(\boldsymbol{\phi}, \boldsymbol{\theta})$ the Hessian matrix for $S_{\mathrm{C}}(\boldsymbol{\phi}, \boldsymbol{\theta})$, the matrix

$$2\widetilde{\sigma_{\mathrm{C}}} H_{\mathrm{C}}^{-1}\left(\widetilde{\boldsymbol{\phi}}_{\mathrm{C}}, \widetilde{\boldsymbol{\theta}}_{\mathrm{C}}\right)$$

can be used as an estimated covariance matrix for $\left(\widetilde{\boldsymbol{\phi}}_{\mathrm{C}}, \widetilde{\boldsymbol{\theta}}_{\mathrm{C}}\right)$.

3.4 Model Checking/Diagnosis Tools for ARMA Models

As in any other version of statistical analysis, it is standard after fitting a time series model to look critically at the quality of that fit, essentially asking "Is the fitted model plausible as a description of what we've seen in the data?" The methodology for doing this examination is based on residuals (in perfect analogy with what is done in ordinary regression analysis). That is, for a fixed set of ARMA parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ^2 the innovations algorithm produces one-step-ahead prediction errors (innovations) (that actually depend only upon $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ and not on σ^2)

$$e_t^{\boldsymbol{\phi},\boldsymbol{\theta}} = y_{t+1} - \hat{y}_{t+1}^{\boldsymbol{\phi},\boldsymbol{\theta}} = y_{t+1} - P_t^{\boldsymbol{\phi},\boldsymbol{\theta}} y_{t+1}$$

and corresponding variances (that additionally depend upon σ^2)

$$\mathrm{E}\left(y_{t+1} - P_t^{\boldsymbol{\phi},\boldsymbol{\theta}} y_{t+1}\right)^2 = v_t^{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2} = \sigma^2 r_t^{\boldsymbol{\phi},\boldsymbol{\theta}}$$

Under the ARMA model (with parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ^2), standardized versions of the prediction errors,

$$w_t^{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2} = \frac{e_t^{\boldsymbol{\phi},\boldsymbol{\theta}}}{\sigma\sqrt{r_t^{\boldsymbol{\phi},\boldsymbol{\theta}}}} \quad ,$$

constitute a white noise sequence with variance 1. So after fitting an ARMA model, one might expect standardized residuals

$$\hat{e}_t^* \equiv w_t^{\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}},\widehat{\sigma^2}} = \frac{e_t^{\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}}}}{\hat{\sigma}\sqrt{r_t^{\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}}}}}$$

to be *approximately* a white noise sequence with variance 1. Tools for examining a time series looking for departures from white noise behavior are then applied to the $\hat{e}_t^*$ as a way of examining the appropriateness of the ARMA model.

For one thing, if the ARMA model is appropriate, the sample autocorrelation function for $\hat{e}_0^*, \hat{e}_1^*, \hat{e}_2^*, \ldots, \hat{e}_{n-1}^*$ ought to be approximately 0 at all lags s bigger than 0. It is thus standard to plot the sample autocorrelation function for the $\hat{e}_t^*$'s (say $\hat{\rho}_n^{\hat{e}^*}(s)$) with limits

$$\pm 2\frac{1}{\sqrt{n}}$$

drawn on the plot, interpreting standardized residuals outside these limits as suggesting dependence at the corresponding lag not adequately accounted for in the modeling.

A related insight is that the expectation that the sample correlations $\hat{\rho}_n^{\hat{e}^*}(1), \hat{\rho}_n^{\hat{e}^*}(2), \ldots, \hat{\rho}_n^{\hat{e}^*}(h)$ are approximately iid normal with mean 0 and standard deviation $1/\sqrt{n}$ translates to an expectation that

$$Q_h \equiv \sum_{s=1}^{h} \left(\sqrt{n}\hat{\rho}_n^{\hat{e}^*}(s)\right)^2 \dot{\sim} \chi_h^2$$

So approximate p-values derived as χ_h^2 right tail probabilities beyond observed values of Q_h might serve as indicators of autocorrelation at a lag of $s \leq h$ not adequately accounted for in modeling. A slight variant of this idea is based on the Ljung-Box statistic

$$Q_h^{\text{LB}} \equiv n(n+2)\sum_{s=1}^{h} \frac{\left(\hat{\rho}_n^{\hat{e}^*}(s)\right)^2}{n-s}$$

for which the χ_h^2 approximation is thought to be better than for Q_h when there is no real departure from white noise. In either case, standard time series software typically produces approximate p-values for some range of values of $h \geq 1$. SAS FS$^{\text{TM}}$ terms plots of the Q_h^{LB} p-values versus h (on a log scale for probability) "white noise test" plots.

Section 1.6 of BDM has a number of other test statistics that can be applied to the series of standardized residuals $\hat{e}_t$ in an effort to identify any clear departures from a white noise model for the standardized innovations.

4 Some Extensions of the ARMA Class of Models

The ARMA models comprise a set of basic building blocks of time series modeling. Their usefulness can be extended substantially by a variety of devices. We consider some of these next.

4.1 ARIMA(p, d, q) Models

It is frequently possible to remove some kinds of obvious trends from time series through (various kinds of) differencing, thereby producing differenced series that can then potentially be modeled as stationary. To begin, we will thus say that a series $\boldsymbol{Y}$ can be described by an ARIMA(p, d, q) (autoregressive integrated moving average model of orders p, d, and q) model if the series $\boldsymbol{Z} = \mathcal{D}^d \boldsymbol{Y}$ is an ARMA(p, q) series, i.e. if $\boldsymbol{Y}$ satisfies

$$\Phi(\mathcal{B}) \mathcal{D}^d \boldsymbol{Y} = \Theta(\mathcal{B}) \boldsymbol{\epsilon}$$

for $\boldsymbol{\epsilon}$ white noise and invertible backshift polynomial operators

$$\Phi(\mathcal{B}) = \mathcal{I} - \phi_1 \mathcal{B} - \phi_2 \mathcal{B}^2 - \cdots - \phi_p \mathcal{B}^p$$

and

$$\Theta(\mathcal{B}) = \mathcal{I} + \theta_1 \mathcal{B} + \theta_2 \mathcal{B}^2 + \cdots + \theta_q \mathcal{B}^q$$

with corresponding polynomials $\phi(z)$ and $\theta(z)$ having no common roots. (The word "integrated" derives from the fact that as values of $\boldsymbol{Z} = \mathcal{D}^d \boldsymbol{Y}$ are derived from values of $\boldsymbol{Y}$ through differencing, values of $\boldsymbol{Y}$ are recovered from values of $\boldsymbol{Z}$ through summing or "integrating.")

Obviously enough, fitting and inference for the ARMA parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ^2 based on the $\boldsymbol{Z}$ series proceeds essentially as in Section 3. One slight adjustment that must be faced is that a realized/observable series

$$\boldsymbol{Y}_n = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

produces an observable vector

$$\boldsymbol{Z}_n = \begin{pmatrix} z_{d+1} \\ z_{d+2} \\ \vdots \\ z_n \end{pmatrix}$$

that is of length only $n - d$ and, of course, $\boldsymbol{Y}_n$ cannot be recovered from $\boldsymbol{Z}_n$ alone. This latter fact might initially cause some practical concern, as one

ultimately typically wants to predict values y_{n+s}, not values z_{n+s}. But, in fact, what needs doing is actually fairly straightforward.

Here we're going to let $\hat{z}_n$ stand for the ARMA(p,q) linear predictor of z_{n+1} based on $\boldsymbol{Z}_n$. Remember that care must then be taken in applying ARMA formulas for predictors and prediction variances from previous sections, since there the observable series begins with index $t = 1$ and here it begins with index $t = d + 1$. We're going to argue that under a natural assumption on the relationship of the differenced series to $\boldsymbol{Y}_d$, the best linear predictor for y_{n+1} is easily written in terms of $\boldsymbol{Y}_d, \boldsymbol{Z}_n$, and $\hat{z}_n$.

Notice that $\boldsymbol{Y}_n$ can be recovered from $\boldsymbol{Y}_d$ and $\boldsymbol{Z}_n$ by a simple recursion. That is, since

$$\mathcal{D}^d = (\mathcal{I} - \mathcal{B})^d = \sum_{j=0}^{d} \binom{d}{j} (-1)^j \mathcal{B}^j$$

one has for $d = 1$

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

so that y_1 and $z_2, \ldots, z_n$ produce $\boldsymbol{Y}_n$, for $d = 2$

$$y_t = z_t + 2y_{t-1} - y_{t-2}$$

so that y_1, y_2 and $z_3, \ldots, z_n$ produce $\boldsymbol{Y}_n$, for $d = 3$

$$y_t = z_t + 3y_{t-1} - 3y_{t-2} + y_{t-3}$$

so that y_1, y_2, y_3 and $z_4, \ldots, z_n$ produce $\boldsymbol{Y}_n$, etc. In general

$$y_t = z_t - \sum_{j=1}^{d} \binom{d}{j} (-1)^j y_{t-j}$$

and knowing $\boldsymbol{Y}_d$ and $\boldsymbol{Z}_n$ is completely equivalent to knowing $\boldsymbol{Y}_n$, and in fact there is a non-singular $n \times n$ matrix $\boldsymbol{C}_n$ such that

$$\boldsymbol{Y}_n = \boldsymbol{C}_n \begin{pmatrix} \boldsymbol{Y}_d \\ \boldsymbol{Z}_n \end{pmatrix}$$

and the (n, n) entry of such a $\boldsymbol{C}_n$ is 1.

So now consider a multivariate Gaussian model for

$$\begin{pmatrix} \boldsymbol{Y}_d \\ \boldsymbol{Z}_{n+1} \end{pmatrix}$$

The ARMA structure gives us an $(n - d + 1)$-dimensional Gaussian model for $\boldsymbol{Z}_{n+1}$. A plausible default assumption is then that the start-up vector $\boldsymbol{Y}_d$ is independent of this vector of d-order differences. Under this assumption, the conditional mean of y_{n+1} given $\boldsymbol{Y}_n$ is

$$\mathrm{E}\left[y_{n+1} | \boldsymbol{Y}_n\right] = \mathrm{E}\left[\boldsymbol{c}_{n+1} \begin{pmatrix} \boldsymbol{Y}_d \\ \boldsymbol{Z}_{n+1} \end{pmatrix} | \boldsymbol{Y}_n\right]$$

for $\boldsymbol{c}_{n+1}$ the last row of $\boldsymbol{C}_{n+1}$. But this is then (since $\boldsymbol{Z}_n$ is a function of $\boldsymbol{Y}_n$)

$$\begin{aligned}\boldsymbol{c}_{n+1}\begin{pmatrix}\boldsymbol{Y}_d\\ \boldsymbol{Z}_n\\ \mathrm{E}\left[z_{n+1}|\boldsymbol{Y}_n\right]\end{pmatrix} &= \boldsymbol{c}_{n+1}\begin{pmatrix}\boldsymbol{Y}_d\\ \boldsymbol{Z}_n\\ \mathrm{E}\left[z_{n+1}|\boldsymbol{Y}_d,\boldsymbol{Z}_n\right]\end{pmatrix}\\ &= \boldsymbol{c}_{n+1}\begin{pmatrix}\boldsymbol{Y}_d\\ \boldsymbol{Z}_n\\ \hat{z}_n\end{pmatrix}\end{aligned}$$

and we see that one simply finds the ARMA predictor for z_{n+1} based on the observed $\boldsymbol{Z}_n$ and uses it in place of z_n in what would be the linear reconstruction of y_{n+1} based on $\boldsymbol{Y}_d$ and $\boldsymbol{Z}_{n+1}$. In fact, since the last entry of the row vector $\boldsymbol{c}_{n+1}$ is 1, the conditional distribution of $y_{n+1}|\boldsymbol{Y}_n$ is Gaussian with this mean and variance that is the ARMA prediction variance for $\hat{z}_n$ (that we note again is not $v_n(1)$ but rather $v_{n-d}(1)$). This line of reasoning then provides sensible (Gaussian model) prediction intervals for y_{n+1} as

$$\boldsymbol{c}_{n+1}\begin{pmatrix}\boldsymbol{Y}_d\\ \boldsymbol{Z}_n\\ \hat{z}_n\end{pmatrix} \pm z\sqrt{v_{n-d}(1)}$$

This development of a predictor (as a conditional mean) and its prediction variance in a Gaussian ARIMA model then says what results when, more generally, one replaces the independence assumption for $\boldsymbol{Y}_d$ and $\boldsymbol{Z}_n$ with an assumption that there is no correlation between any entry of the first and an entry of the second and looks for best linear predictors.

4.2 SARIMA$(p,d,q)\times(P,D,Q)_s$ Models

Particularly in economic forecasting contexts, one often needs to do seasonal differencing in order to remove more or less obviously regular patterns in a time series. For example, with quarterly data, I might want to apply the operator

$$\mathcal{D}_4 = \mathcal{I} - \mathcal{B}^4$$

to a raw time series $\boldsymbol{Y}$ before trying to model $\boldsymbol{Z} = \mathcal{D}_4\boldsymbol{Y}$ as ARMA(p,q). The standard generalization of this idea is the so-called SARIMA (seasonal ARIMA) class of models.

We'll say that $\boldsymbol{Y}$ is described by a SARIMA$(p,d,q)\times(P,D,Q)_s$ model provided

$$\boldsymbol{Z} = (\mathcal{I}-\mathcal{B})^d\left(\mathcal{I}-\mathcal{B}^s\right)^D\boldsymbol{Y}$$

has a representation in terms of a causal ARMA model defined by the equation

$$\Phi(\mathcal{B})\,\Phi_s(\mathcal{B}^s)\,\boldsymbol{Z} = \Theta(\mathcal{B})\,\Theta_s(\mathcal{B}^s)\,\boldsymbol{\epsilon} \tag{37}$$

where $\boldsymbol{\epsilon}$ is white noise,

$$\begin{aligned}\Phi(\mathcal{B}) &= \mathcal{I}-\phi_1\mathcal{B}-\phi_2\mathcal{B}^2-\cdots-\phi_p\mathcal{B}^p\\ \Phi_s(\mathcal{B}^s) &= \mathcal{I}-\phi_{s,1}\mathcal{B}^s-\phi_{s,2}\mathcal{B}^{2s}-\cdots-\phi_{s,P}\mathcal{B}^{Ps}\\ \Theta(\mathcal{B}) &= \mathcal{I}+\theta_1\mathcal{B}+\theta_2\mathcal{B}^2+\cdots+\theta_q\mathcal{B}^q\end{aligned}$$

and

$$\Theta_s\left(\mathcal{B}^s\right) = \mathcal{I} + \theta_{s,1}\mathcal{B}^s + \theta_{s,2}\mathcal{B}^{2s} + \cdots + \theta_{s,Q}\mathcal{B}^{Qs}$$

Clearly, the operators

$$\Phi^*\left(\mathcal{B}\right) \equiv \Phi\left(\mathcal{B}\right)\Phi_s\left(\mathcal{B}^s\right) \quad \text{and} \quad \Theta^*\left(\mathcal{B}\right) \equiv \Theta\left(\mathcal{B}\right)\Theta_s\left(\mathcal{B}^s\right)$$

are backshift polynomial operators of respective orders $p + sP$ and $q + sQ$, and the basic SARIMA equation can be written as

$$\Phi^*\left(\mathcal{B}\right)\boldsymbol{Z} = \Theta^*\left(\mathcal{B}\right)\boldsymbol{\epsilon}$$

that obviously specifies a special ARIMA model. Once one has written the model equation this way, it is more or less clear how to operate. One must fit parameters $\boldsymbol{\phi}^*, \boldsymbol{\theta}^*$, and σ^2 (of total dimension $p + P + q + Q + 1$) using $\boldsymbol{Z}_n$ (that is of length $n - (d + sD)$). (While $\Phi^*\left(\mathcal{B}\right)$ is of order $p + sP$ and in some cases there are this many coefficients specifying $\Phi^*\left(\mathcal{B}\right)$, there are only $p + P$ parameters involved in defining those coefficients. While $\Theta^*\left(\mathcal{B}\right)$ is of order $q + sQ$, there are only $q + Q$ parameters involved in defining the coefficients for $\Theta^*\left(\mathcal{B}\right)$.) Prediction for z_{n+s} proceeds from knowing how to do ARMA prediction, and then (ARIMA) prediction for y_{n+s} follows from an assumption that $\boldsymbol{Y}_{d+sD}$ is uncorrelated with $\boldsymbol{Z}_n$.

The restriction to causal forms (37) is the restriction to cases where $\Phi^*\left(\mathcal{B}\right)$ is invertible, is the restriction to forms where the polynomial corresponding to $\Phi^*\left(\mathcal{B}\right)$ has no roots inside the unit circle. This, in turn, is the restriction to parameter sets where polynomials corresponding to both $\Phi\left(\mathcal{B}\right)$ and $\Phi_s\left(\mathcal{B}^s\right)$ have no roots inside the unit circle.

The form (37) can be motivated as follows. Suppose, for example, that $s = 4$ (this would be sensible where quarterly economic data are involved). Let

$$\boldsymbol{Z}^1 = \begin{pmatrix} \vdots \\ z_{-3} \\ z_1 \\ z_5 \\ \vdots \end{pmatrix}, \boldsymbol{Z}^2 = \begin{pmatrix} \vdots \\ z_{-2} \\ z_2 \\ z_6 \\ \vdots \end{pmatrix}, \boldsymbol{Z}^3 = \begin{pmatrix} \vdots \\ z_{-1} \\ z_3 \\ z_7 \\ \vdots \end{pmatrix}, \text{ and } \boldsymbol{Z}^4 = \begin{pmatrix} \vdots \\ z_0 \\ z_4 \\ z_8 \\ \vdots \end{pmatrix}$$

and consider the possibility that for $\boldsymbol{U}$ white noise and

$$\boldsymbol{U}^1 = \begin{pmatrix} \vdots \\ u_{-3} \\ u_1 \\ u_5 \\ \vdots \end{pmatrix}, \boldsymbol{U}^2 = \begin{pmatrix} \vdots \\ u_{-2} \\ u_2 \\ u_6 \\ \vdots \end{pmatrix}, \boldsymbol{U}^3 = \begin{pmatrix} \vdots \\ u_{-1} \\ u_3 \\ u_7 \\ \vdots \end{pmatrix}, \text{ and } \boldsymbol{U}^4 = \begin{pmatrix} \vdots \\ u_0 \\ u_4 \\ u_8 \\ \vdots \end{pmatrix}$$

there are sets of P coefficients $\boldsymbol{\phi}_4$ and Q coefficients $\boldsymbol{\theta}_4$ and corresponding order P and Q backshift polynomials $\boldsymbol{\Phi}_4\left(\mathcal{B}\right)$ and $\boldsymbol{\Theta}_4\left(\mathcal{B}\right)$ for which

$$\boldsymbol{\Phi}_4\left(\mathcal{B}\right)\boldsymbol{Z}^j = \boldsymbol{\Theta}_4\left(\mathcal{B}\right)\boldsymbol{U}^j \quad \text{for } j = 1, 2, 3, 4$$

This means that the $\boldsymbol{Z}^j$ are uncorrelated, all governed by the same form of ARMA(P,Q) model. These equations imply that

$$z_t - \phi_{4,1} z_{t-4} - \phi_{4,2} z_{t-8} - \cdots - \phi_{4,P} z_{t-4P} = u_t + \theta_{4,1} u_{t-4} + \theta_{4,2} u_{t-8} + \cdots + \theta_{4,Q} u_{t-4Q} \ \forall t$$

which in other notation means that

$$\boldsymbol{\Phi}_4\left(\mathcal{B}^4\right)\boldsymbol{Z} = \boldsymbol{\Theta}_4\left(\mathcal{B}^4\right)\boldsymbol{U} \tag{38}$$

Consider the $P = 1$ and $Q = 0$ version of this possibility. In this case, the autocorrelation function for $\boldsymbol{Z}$ is 0 at lags that are not multiples of 4, and for integer k, $\rho(4k) = \phi_{4,1}^{|k|}$. On the other hand, for $P = 0$ and $Q = 1$, the autocorrelation is 0 except at lag $s = 4$.

Now the fact that if $\boldsymbol{U}$ is white noise the $\boldsymbol{Z}^j$ are uncorrelated (independent in Gaussian cases) is usually intuitively not completely satisfying. But, if $\boldsymbol{U}$ were MA(q) for $q < 4$ then the $\boldsymbol{Z}^j$ would have the same distributions as when $\boldsymbol{U}$ is white noise, but would not be uncorrelated. Or, if $\boldsymbol{U}$ were AR with small coefficients $\boldsymbol{\phi}$, the model for a $\boldsymbol{Z}^j$ might be nearly the same as for $\boldsymbol{U}$ white noise, but again successive z_t's would not be uncorrelated.

So one is led to consider generalizing this development by replacing a white noise assumption on $\boldsymbol{U}$ with an ARMA(p,q) assumption. That is, for invertible $\boldsymbol{\Phi}(\mathcal{B})$ and $\boldsymbol{\Theta}(\mathcal{B})$ and white noise $\boldsymbol{\epsilon}$, as an alternative to relationship (38) we might consider a model equation

$$\boldsymbol{\Phi}_4\left(\mathcal{B}^4\right)\boldsymbol{Z} = \boldsymbol{\Theta}_4\left(\mathcal{B}^4\right)\boldsymbol{\Phi}^{-1}(\mathcal{B})\,\boldsymbol{\Theta}(\mathcal{B})\,\boldsymbol{\epsilon}$$

or applying $\boldsymbol{\Phi}(\mathcal{B})$ to both sides

$$\begin{aligned}\boldsymbol{\Phi}(\mathcal{B})\,\boldsymbol{\Phi}_4\left(\mathcal{B}^4\right)\boldsymbol{Z} &= \boldsymbol{\Phi}(\mathcal{B})\,\boldsymbol{\Theta}_4\left(\mathcal{B}^4\right)\boldsymbol{\Phi}^{-1}(\mathcal{B})\,\boldsymbol{\Theta}(\mathcal{B})\,\boldsymbol{\epsilon} \\ &= \boldsymbol{\Theta}(\mathcal{B})\,\boldsymbol{\Theta}_4\left(\mathcal{B}^4\right)\boldsymbol{\epsilon}\end{aligned} \tag{39}$$

This is an $s = 4$ version of the general SARIMA equation (37). In the present situation we expect that replacing a white noise model for $\boldsymbol{U}$ with an ARMA model to produce a model for $\boldsymbol{Z}$ in which there are relatively big autocorrelations around lags that are multiples of 4, but that also allows for *some* correlation at other lags. In general, we expect a SARIMA model to have associated autocorrelations that are "biggish" around lags that are multiples of s, but that can be non-negligible at other lags as well.

For sake of concreteness and illustration, consider the SARIMA$(1,1,1) \times (0,1,1)_4$ version of relationship (39). This (for $\boldsymbol{Z} = (\mathcal{I} - \mathcal{B})\left(\mathcal{I} - \mathcal{B}^4\right)\boldsymbol{Y} = \left(\mathcal{I} - \mathcal{B} - \mathcal{B}^4 + \mathcal{B}^5\right)\boldsymbol{Y}$) is

$$\boldsymbol{Z} = \left(\mathcal{I} + \theta_{4,1}\mathcal{B}^4\right)\boldsymbol{U}$$

where $\boldsymbol{U}$ is ARMA$(1,1)$. That is,

$$\boldsymbol{Z} = \left(\mathcal{I} + \theta_{4,1}\mathcal{B}^4\right)(\mathcal{I} - \phi_1\mathcal{B})^{-1}(\mathcal{I} + \theta_1\mathcal{B})\,\boldsymbol{\epsilon}$$

for appropriate constants $\theta_{4,1}, \phi_1$, and θ_1 and white noise $\boldsymbol{\epsilon}$. But this is

$$\begin{aligned}(\mathcal{I} - \phi_1\mathcal{B})\,\boldsymbol{Z} &= \left(\mathcal{I} + \theta_{4,1}\mathcal{B}^4\right)(\mathcal{I} + \theta_1\mathcal{B})\,\boldsymbol{\epsilon} \\ &= \left(\mathcal{I} + \theta_1\mathcal{B} + \theta_{4,1}\mathcal{B}^4 + \theta_1\theta_{4,1}\mathcal{B}^5\right)\boldsymbol{\epsilon}\end{aligned}$$

and it is now evident that this is a special ARMA$(1,5)$ model for $\boldsymbol{Z}$ (that is itself a very special kind of 5th order backshift polynomial function of $\boldsymbol{Y}$), where three of the potentially $p + q + Ps + Qs = 1 + 1 + 0 + 4 \cdot 1 = 6$ ARMA coefficients are structurally 0, and the four that are not are functions of only three parameters, ϕ_1, θ_1, and $\theta_{4,1}$. So then, how to proceed is more or less obvious. Upon estimating the parameters (by maximum likelihood or some other method) prediction here works exactly as in any ARIMA model.

This discussion of the fact that SARIMA models are special ARIMA models (under alternative parameterizations) brings up a related matter, that of **subset models**. That is, in "ordinary" ARIMA(p,q) modeling, it is possible to consider purposely setting to 0 particular coefficients in the defining polynomial backshift operators $\Phi(\mathcal{B})$ and $\Theta(\mathcal{B})$. The resulting model has fewer than $p + q + 1$ parameters that can be estimated by maximum likelihood or other methods. And once this is done, prediction can proceed as for any ARIMA model. The `SAS FS`$^{\text{TM}}$ software allows the specification and use of such models, but as far as I can tell, the `JMP`$^{\text{TM}}$ software does not.

4.2.1 A Bit About "Intercept" Terms and Differencing in ARIMA (and SARIMA) Modeling

The purpose of various kinds of differencing of a time series, $\boldsymbol{Y}$, is the removal of trend and corresponding reduction to a differenced series for which a stationary model is appropriate. One of the options that standard time series software usually provides is the use of an "intercept" in ARIMA modeling. Where $\boldsymbol{Z}$ is derived from $\boldsymbol{Y}$ by some form of differencing, this is ARMA modeling of not $\boldsymbol{Z}$ but rather $\boldsymbol{Z} - \mu\mathbf{1}$ for a real parameter μ. That is, this is use of a model specified by

$$\Phi(\mathcal{B})\,(\boldsymbol{Z} - \mu\mathbf{1}) = \Theta(\mathcal{B})\,\boldsymbol{\epsilon}$$

Where (as is common) $\Phi(\mathcal{B})$ has an inverse, this is

$$\boldsymbol{Z} = \mu\mathbf{1} + \Phi(\mathcal{B})^{-1}\,\Theta(\mathcal{B})\,\boldsymbol{\epsilon}$$

and $\mathrm{E}\boldsymbol{Z} = \mu\mathbf{1}$.

In many contexts, this is problematic if $\mu \neq 0$. For $\mathcal{D}^r$ some difference operator, $\mathrm{E}\boldsymbol{Z} = \mu\mathbf{1}$ implies that $\mathrm{E}y_t$ is an r-degree polynomial in t with leading coefficient $\mu \neq 0$. That means that for large t, $\mathrm{E}y_t$ is of order t^r, "exploding" to $\pm\infty$ (depending upon the sign of μ). Typically this is undesirable, particularly where one needs to make forecasts $\hat{y}_{n+s}$ for large s (and for which $\mathrm{E}y_t$ will dominate the computations). Note that if $\mathrm{E}\boldsymbol{Z} = \mu\mathbf{1}$ it's the case that $\mathrm{E}\mathcal{D}\boldsymbol{Z} = \mathbf{0}$. So "another" differencing applied to a $\boldsymbol{Z}$ that needs an intercept in modeling, produces a mean 0 series.

But this last observation is not an indication that one should go wild with differencing. Differencing reduces the length of a time series available for model fitting, and can actually increase the complexity of a model needed to describe a situation. To see this, consider a situation where $\boldsymbol{Y}$ (that could have been produced by differencing some original series) is ARMA(p,q). That is, for white noise $\boldsymbol{\epsilon}$,

$$\Phi(\mathcal{B})\boldsymbol{Y} = \Theta(\mathcal{B})\boldsymbol{\epsilon}$$

for orders p and q polynomial backshift operators $\Phi(\mathcal{B})$ and $\Theta(\mathcal{B})$. Suppose that $\boldsymbol{Y}$ is differenced. This produces $\mathcal{D}\boldsymbol{Y}$ that solves

$$\Phi(\mathcal{B})\mathcal{D}\boldsymbol{Y} = \mathcal{D}\Theta(\mathcal{B})\boldsymbol{\epsilon}$$

Thus (provided $\Phi(\mathcal{B})$ is invertible) $\mathcal{D}\boldsymbol{Y}$ has a causal ARMA$(p, q+1)$ representation, where (by virtue of the fact that the polynomial corresponding to $\mathcal{D}$ has a root at 1) the model is *not* invertible. This has made the modeling more complicated (in the move from ARMA(p,q) to ARMA$(p,q+1)$).

So, one wants to difference only enough to remove a trend. It's possible to "over-difference" and ultimately make ARIMA modeling less than simple and interpretable.

4.3 Regression Models With ARMA Errors

For $\boldsymbol{w}_n$ a vector or matrix of observed covariates/predictors (values of variables that might be related to $\boldsymbol{Y}_n$) we might wish to model as

$$\boldsymbol{Y}_n = \boldsymbol{g}_n(\boldsymbol{w}_n, \boldsymbol{\beta}) + \boldsymbol{Z}_n$$

for $\boldsymbol{Z}_n$ consisting of entries 1 through n of an ARIMA(p,d,q) series $\boldsymbol{Z}$, $\boldsymbol{g}_n$ is a known function mapping to $\Re^n$, and $\boldsymbol{\beta}$ is k-vector of parameters. Probably the most important possibilities here include those where $\boldsymbol{w}_n$ includes some elements of one or more other time series (through time at most n) that one hopes "lead" the $\boldsymbol{Y}$ series and $\boldsymbol{w}_t$ contains all the values in $\boldsymbol{w}_{t-1}$, the function $\boldsymbol{g}_n$ is of the form

$$\boldsymbol{g}_n(\boldsymbol{w}_n, \boldsymbol{\beta}) = \begin{pmatrix} g_1(\boldsymbol{w}_1, \boldsymbol{\beta}) \\ g_2(\boldsymbol{w}_2, \boldsymbol{\beta}) \\ \\ \\ g_n(\boldsymbol{w}_n, \boldsymbol{\beta}) \end{pmatrix}$$

for real-valued functions g_t that are (across t) related in natural ways, and the parameter vector has the same meaning/role for all t.

At least the Gaussian version of model fitting here is more or less obvious. Where $\boldsymbol{Z}$ is ARMA,

$$\boldsymbol{Y}_n \sim \text{MVN}_n\left(\left(\boldsymbol{g}_n(\boldsymbol{w}_n, \boldsymbol{\beta})\right), \boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}\right)$$

(for $\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2$ the ARMA parameters and $\boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}$ the corresponding $n \times n$ covariance matrix) and the likelihood function has the form

$$\begin{aligned} f\left(\boldsymbol{y}_n | \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2\right) &= \frac{1}{\sqrt{(2\pi)^n \left|\det \boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}\right|}} \\ &\quad \times \exp\left(-\frac{1}{2}\left(\boldsymbol{y}_n - \left(\boldsymbol{g}_n\left(\boldsymbol{w}_n, \boldsymbol{\beta}\right)\right)\right)' \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}\left(\boldsymbol{y}_n - \left(\boldsymbol{g}_n\left(\boldsymbol{w}_n, \boldsymbol{\beta}\right)\right)\right)\right) \end{aligned}$$

a function of $m + p + q + 1$ real parameters. This can be used to guide inference for the model parameters, leading to maximum likelihood estimates and approximate confidence limits derived from the estimated covariance matrix (itself derived from the Hessian of the logarithm of this evaluated at the maximum likelihood estimates).

When all entries of $\boldsymbol{w}_{n+s}$ are available at time n, one can make use of the multivariate normal distribution of

$$\begin{pmatrix} \boldsymbol{y}_n \\ y_{n+s} \end{pmatrix}$$

that has mean

$$\begin{pmatrix} \boldsymbol{g}_n\left(\boldsymbol{w}_n, \boldsymbol{\beta}\right) \\ g_{n+s}\left(\boldsymbol{w}_{n+s}, \boldsymbol{\beta}\right) \end{pmatrix}$$

and for $\boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}$ as above has covariance matrix

$$\begin{pmatrix} \underset{n\times n}{\boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}} & \begin{pmatrix} \gamma_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}\left(n+s-1\right) \\ \vdots \\ \gamma_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}\left(s\right) \end{pmatrix} \\ \left(\gamma_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}\left(n+s-1\right), \ldots, \gamma_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}\left(s\right)\right) & \sigma^2 \end{pmatrix}$$

to find the conditional mean of y_{n+s} given $\boldsymbol{y}_n$, that is the best linear predictor for y_{n+s} even without the Gaussian assumption. (Of course, in practice, one will have not $\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2$ but rather estimates $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \widehat{\sigma^2}$, and the prediction will be only some estimated best linear prediction.) And the Gaussian conditional variance for y_{n+s} given $\boldsymbol{y}_n$ provides a prediction variance and prediction limits as well.

All that changes in this story when $\boldsymbol{Z}$ is ARIMA is that one writes

$$\left(\boldsymbol{Y}_n - \boldsymbol{g}_n\left(\boldsymbol{w}_n, \boldsymbol{\beta}\right)\right) = \boldsymbol{Z}_n$$

and then

$$\mathcal{D}^d\left(\boldsymbol{Y}_n - \boldsymbol{g}_n\left(\boldsymbol{w}_n, \boldsymbol{\beta}\right)\right) = \mathcal{D}^d \boldsymbol{Z}_n$$

and carries out the above program with

$$\boldsymbol{Y}^*_n = \mathcal{D}^d \boldsymbol{Y}_n, \boldsymbol{g}^*_n\left(\boldsymbol{w}_n, \boldsymbol{\beta}\right) = \mathcal{D}^d \boldsymbol{g}_n\left(\boldsymbol{w}_n, \boldsymbol{\beta}\right), \quad \text{and} \quad \boldsymbol{Z}^*_n = \mathcal{D}^d \boldsymbol{Z}_n$$

where now $\boldsymbol{Z}^*_n$ consists of $n - d$ elements of the ARMA series $\mathcal{D}^d \boldsymbol{Z}$. Notice that, fairly obviously, the differenced series $\boldsymbol{Y}^*_n$ has mean $\boldsymbol{g}^*_n\left(\boldsymbol{w}_n, \boldsymbol{\beta}\right)$ under this modeling.

4.3.1 Parametric Trends

We proceed to illustrate the usefulness of this formalism in a number of situations. To begin, note that for $g(t|\boldsymbol{\beta})$ a parametric function of t, the choice of $\boldsymbol{w}_n = (1, 2, \ldots, n)$

$$g_n(\boldsymbol{w}_n|\boldsymbol{\beta}) = \begin{pmatrix} g(1|\boldsymbol{\beta}) \\ \vdots \\ g(n|\boldsymbol{\beta}) \end{pmatrix}$$

provides a model for $\boldsymbol{Y}$ that is

$$\textit{parametric trend} + \textit{ARMA noise}$$

(For example, $g(t|\boldsymbol{\beta})$ could be a polynomial of order $m-1$ and the entries of $\boldsymbol{\beta}$ the coefficients of that polynomial.)

4.3.2 "Interventions"

As a second example, consider models with an "intervention"/mean shift at time t_0. If $\boldsymbol{w}$ is such that $w_t = 0$ for $t < t_0$ and $w_t = 1$ for $t \geq t_0$ then an ARMA(p, q) model for $\boldsymbol{Y}$ with an "intervention"/mean shift at time t_0 uses the MVN$_n$ distribution of $\boldsymbol{Y}_n$ with mean $\beta\boldsymbol{w}_n$ and $n \times n$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}$.

An ARIMA(p, d, q) model with mean shift for $\boldsymbol{Y}$ means that

$$\boldsymbol{Z} = \mathcal{D}^d(\boldsymbol{Y} - \beta\boldsymbol{w})$$

is (mean 0) ARMA(p, q). Note for example that with $d = 1$ this prescription makes the differenced $\boldsymbol{Y}$ series have mean

$$\mathrm{E}\mathcal{D}\boldsymbol{Y} = \beta\mathcal{D}\boldsymbol{w}$$

which is a series that is 0 except at time t_0, where it is β. So for fitting the ARMA$(p, 1, q)$ model based on $\boldsymbol{Y}_n$, one uses the MVN$_{n-1}$ distribution of $\mathcal{D}\boldsymbol{Y}_n$ with mean

$$\boldsymbol{g}_n(\boldsymbol{w}_n, \beta) = \begin{pmatrix} \underset{(t_0-2)\times 1}{\mathbf{0}} \\ \beta \\ \underset{(n-t_0)\times 1}{\mathbf{0}} \end{pmatrix}$$

and an $(n-1) \times (n-1)$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}$.

Related to these examples are models with a "pulse intervention/event" at time t_0 (essentially representing an outlier at this period). That is, with $\boldsymbol{w}$ as above (with $w_t = 0$ for $t < t_0$ and $w_t = 1$ for $t \geq t_0$), $\mathcal{D}\boldsymbol{w}$ is a unit pulse at time t_0. An ARMA(p, q) model for $\boldsymbol{Y}$ with a pulse of β at time t_0 uses the MVN$_n$ distribution of $\boldsymbol{Y}_n$ with mean $\beta\mathcal{D}\boldsymbol{w}_n$ and $n \times n$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}$.

An ARIMA(p, d, q) model for $\boldsymbol{Y}$ with a pulse of β at t_0 makes

$$\boldsymbol{Z} = \mathcal{D}^d(\boldsymbol{Y} - \beta\mathcal{D}\boldsymbol{w})$$

ARMA(p,q). In the $d=1$ case, this implies that the mean of the differenced $\boldsymbol{Y}$ series is

$$\mathrm{E}\mathcal{D}\boldsymbol{Y} = \beta\mathcal{D}^2\boldsymbol{w}$$

which is a series that is 0 except at time t_0, where it is β, and at time t_0+1 where it is $-\beta$. So for fitting the ARIMA$(p,1,q)$ model based on $\boldsymbol{Y}_n$, one uses the MVN_{n-1} distribution of $\mathcal{D}\boldsymbol{Y}_n$ with mean

$$\boldsymbol{g}_n(\boldsymbol{w}_n,\beta) = \begin{pmatrix} \underset{(t_0-2)\times 1}{\mathbf{0}} \\ \beta \\ -\beta \\ \underset{(n-1-t_0)\times 1}{\mathbf{0}} \end{pmatrix}$$

and $(n-1)\times(n-1)$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2}$.

The timing of level shifts and/or pulses in a time series analysis is rarely something that can be specified in any but empirical terms. One can sometimes look back over a plot of the values y_t versus t (or at a plot of ARMA or ARIMA residuals $\hat{e}_t^*$ against t) and see that a level shift or designation of one or more values as outliers will be needed to adequately describe a situation. But using a future level shift or pulse in forecasting is not at all common, and would require the specification of both t_0 and β in advance of the occurrence of these events.

4.3.3 "Exogenous Variables"/Covariates and "Transfer Function" Models

Now consider cases of the regression framework of this section where a covariate series $\boldsymbol{x}$ is involved. Suppose then that for some $r \geq 0$ (a "time delay" or "dead time")

$$\Lambda(\mathcal{B}) = \mathcal{B}^r \sum_{j=0}^{m-1} \beta_j \mathcal{B}^j$$

is $\mathcal{B}^r$ "times" a backshift polynomial of order $m-1$, and set

$$g_t(\boldsymbol{w}_t,\boldsymbol{\beta}) = (\Lambda(\mathcal{B})\boldsymbol{x})_t = \sum_{j=r}^{r+m-1} \beta_j x_{t-j}$$

for $\boldsymbol{\beta} = \left(\beta_r, \beta_{r+1}, \ldots, \beta_{r+m-1}\right)$ and $\boldsymbol{w}_t = (x_{1-m-r}, \ldots, x_{t-r})$ producing

$$\boldsymbol{g}_n(\boldsymbol{w}_n,\boldsymbol{\beta}) = \begin{pmatrix} \sum_{j=r}^{r+m-1} \beta_j x_{1-j} \\ \vdots \\ \sum_{j=r}^{r+m-1} \beta_j x_{n-j} \end{pmatrix}$$

Then (depending upon how far into the past one has available values of the $\boldsymbol{x}$ series) a multivariate normal distribution for some final part of $\boldsymbol{Y}_n$ can be used in fitting the coefficients of an ARMA or ARIMA model for $\boldsymbol{Y}$. Assuming

that values of x_t through time n are available, means $g_t(\boldsymbol{w}_t, \boldsymbol{\beta}) = \mathrm{E}y_t$ through time $t = n + r$ are available, and so too are Gaussian conditional means/linear forecasts of y_t and Gaussian prediction limits.

It is worth noting at this point that in economic forecasting contexts, it's common for external series $\boldsymbol{x}_n$ available at time n to come with forecasts $\hat{x}_{n+1}, \hat{x}_{n+2}, \hat{x}_{n+3}, \ldots$ (produced from unspecified sources and considerations). In such cases, these are often used in place of honest observed values x_{n+s} in the form

$$\sum_{j=r}^{r+m-1} \beta_j x_{t-j}$$

to produce approximate values of $g_t(\boldsymbol{w}_t, \boldsymbol{\beta})$ for $t > n + r$ in order to forecast beyond time $t = n + r$.

It appears that `SAS FS`$^{\mathrm{TM}}$ is willing to somehow make forecasts beyond period $t = n + r$ even in the absence of input forecasts $\hat{x}_{n+1}, \hat{x}_{n+2}, \hat{x}_{n+3}, \ldots$. I honestly don't know what the program is doing in those cases. My best guess is that $\hat{x}_{n+1}, \hat{x}_{n+2}, \hat{x}_{n+3}, \ldots$ are all set to the last observed value, x_n.

A model for $\boldsymbol{Y}$ that says that

$$y_t = \sum_{j=r}^{r+m-1} \beta_j x_{t-j} + z_t \quad \forall t$$

where z_t is mean 0 ARMA(p, q) noise can be written in other terms as

$$\boldsymbol{Y} = \Lambda(\mathcal{B})\,\boldsymbol{x} + \boldsymbol{Z}$$

where $\boldsymbol{Z}$ solves

$$\Phi(\mathcal{B})\,\boldsymbol{Z} = \Theta(\mathcal{B})\,\boldsymbol{\epsilon}$$

for $\boldsymbol{\epsilon}$ white noise. That is, $\boldsymbol{Y}$ solves

$$\Phi(\mathcal{B})\,(\boldsymbol{Y} - \Lambda(\mathcal{B})\,\boldsymbol{x}) = \Theta(\mathcal{B})\,\boldsymbol{\epsilon}$$

Or, for example, an ARIMA$(p, 1, q)$ model for $\boldsymbol{Y}$ with mean function $\Lambda(\mathcal{B})\,\boldsymbol{x}$ makes

$$\mathcal{D}\,(\boldsymbol{Y} - \Lambda(\mathcal{B})\,\boldsymbol{x})$$

ARMA(p, q) and the mean of the differenced $\boldsymbol{Y}$ series is

$$\mathrm{E}\mathcal{D}\boldsymbol{Y} = \Lambda(\mathcal{B})\,\mathcal{D}\boldsymbol{x}$$

A generalization of this development that is sometimes put forward as an effective modeling tool is the use of not "backshift polynomials" but rather "rational functions in the backshift operator." The idea here is to (with $\Lambda(\mathcal{B})$ as above) consider time-invariant linear operators of the form

$$\Omega(\mathcal{B})^{-1}\,\Lambda(\mathcal{B})$$

for

$$\Omega(\mathcal{B}) = \mathcal{I} - \alpha_1\mathcal{B} - \alpha_2\mathcal{B}^2 - \cdots - \alpha_l\mathcal{B}^l$$

as "transfer functions," mapping input $\boldsymbol{x}$ series to mean functions for $\boldsymbol{Y}$. In order to get some insight into what this promises to do, consider the case where $r = 2, m = 3$ and $l = 1$ where it is easy to see in quite explicit terms what this operator looks like. That is, for $|\alpha_1| < 1$

$$\begin{aligned}
\Omega(\mathcal{B})^{-1}\Lambda(\mathcal{B}) &= (\mathcal{I} - \alpha_1\mathcal{B})^{-1}\left(\mathcal{B}^2\left(\beta_0\mathcal{B}^0 + \beta_1\mathcal{B}^1 + \beta_2\mathcal{B}^2\right)\right) \\
&= \left(\sum_{s=0}^{\infty}\alpha_1^s\mathcal{B}^s\right)\left(\beta_0\mathcal{B}^2 + \beta_1\mathcal{B}^3 + \beta_2\mathcal{B}^4\right) \\
&= \beta_0\mathcal{B}^2 + (\alpha_1\beta_0 + \beta_1)\mathcal{B}^3 + \left(\alpha_1^2\beta_0 + \alpha_1\beta_1 + \beta_2\right)\mathcal{B}^4 \\
&\quad + \sum_{s=5}^{\infty}\left(\alpha_1^2\beta_0 + \alpha_1\beta_1 + \beta_2\right)\alpha_1^{s-4}\mathcal{B}^s
\end{aligned}$$

That is, this time-invariant linear operator $\Omega(\mathcal{B})^{-1}\Lambda(\mathcal{B})$ has impulse response function with values

$$\psi_s = \begin{cases} 0 & \text{for } s < 2 \\ \beta_0 & \text{for } s = 2 \\ \alpha_1\beta_0 + \beta_1 & \text{for } s = 3 \\ \left(\alpha_1^2\beta_0 + \alpha_1\beta_1 + \beta_2\right)\alpha_1^{s-4} & \text{for } s \geq 4 \end{cases}$$

Here, $r = 2$ governs the delay, $m = 3$ governs how long the coefficients of $\Lambda(\mathcal{B})$ directly impact the character of the impulse response (beyond $s = m$ the character of the impulse response is the exponential decay character of that of $\Omega(\mathcal{B})^{-1}$) and for $s \leq m$ the coefficients in $\Lambda(\mathcal{B})$ and $\Omega(\mathcal{B})^{-1}$ interact in a patterned way to give the coefficients for $\Omega(\mathcal{B})^{-1}\Lambda(\mathcal{B})$. These things are not special to this case, but are in general qualitatively how the two backshift polynomials combine to produce this rational function of the backshift operator.

Clearly then,

$$\boldsymbol{Y} = \Omega(\mathcal{B})^{-1}\Lambda(\mathcal{B})\boldsymbol{x} + \boldsymbol{Z} \tag{40}$$

where $\boldsymbol{Z}$ solves

$$\Phi(\mathcal{B})\boldsymbol{Z} = \Theta(\mathcal{B})\boldsymbol{\epsilon}$$

for $\boldsymbol{\epsilon}$ white noise is a model for $\boldsymbol{Y}$ with mean $\Omega(\mathcal{B})^{-1}\Lambda(\mathcal{B})\boldsymbol{x}$ and ARMA deviations from that mean function. Of course, rewriting slightly, in this case $\boldsymbol{Y}$ solves

$$\Phi(\mathcal{B})\left(\boldsymbol{Y} - \Omega(\mathcal{B})^{-1}\Lambda(\mathcal{B})\boldsymbol{x}\right) = \Theta(\mathcal{B})\boldsymbol{\epsilon} \tag{41}$$

Or, for example, an ARIMA$(p, 1, q)$ model for $\boldsymbol{Y}$ with mean function $\Omega(\mathcal{B})^{-1}\Lambda(\mathcal{B})\boldsymbol{x}$ makes

$$\mathcal{D}\left(\boldsymbol{Y} - \Omega(\mathcal{B})^{-1}\Lambda(\mathcal{B})\boldsymbol{x}\right)$$

ARMA(p, q) and the mean of the differenced $\boldsymbol{Y}$ series is

$$\mathrm{E}\mathcal{D}\boldsymbol{Y} = \Omega(\mathcal{B})^{-1}\Lambda(\mathcal{B})\mathcal{D}\boldsymbol{x}$$

Gaussian-based inference for a "transfer function model" like models (40) and (41) is again more or less "obvious," as is subsequent prediction/forecasting using fitted model parameters as "truth," provided one has an adequate set of values from $\boldsymbol{x}$ to support calculation of the model's mean for y_{n+s}. That is, a representation like (40) or (41) provides a mean for $\boldsymbol{Y}_n$ that is a function of the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2$, and a covariance matrix that is a function of $\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2$. Adoption of a Gaussian assumption provides a likelihood function for these ($l+m+p+q+1$ univariate) parameters that can guide inference for them. Forecasting then proceeds as for all ARMA models.

4.3.4 Sums of the Above Forms for E$\boldsymbol{Y}$

It is, of course, possible that one might want to include more than one of the basic cases of $\boldsymbol{g}_n(\boldsymbol{w}_n, \boldsymbol{\beta})$ laid out above in a single time series model. There could be in a single application reason to use a parametric trend, multiple intervention events, and multiple covariates. There is really no conceptual problem with handling these in a single additive form for $\mathrm{E}\boldsymbol{Y}_n$. That is, for k instances of the basic structure under discussion,

$$\boldsymbol{g}_n^i\left(\boldsymbol{w}_n^i, \boldsymbol{\beta}^i\right) \text{ for } i = 1, 2, \ldots, k$$

One can employ a mean function of the form

$$\boldsymbol{g}_n(\boldsymbol{w}_n, \boldsymbol{\beta}) = \sum_{i=1}^{k} \boldsymbol{g}_n^i\left(\boldsymbol{w}_n^i, \boldsymbol{\beta}^i\right)$$

and (with $\sum_{i=1}^{k} m_i$ univariate parameters β_j^i) handle for example multiple outliers, multiple covariate series, etc. in a single regression model with ARMA (or ARIMA) noise.

4.3.5 Regression or Multivariate Time Series Analysis?

It should be emphasized before going further that the discussion here has treated the covariate as "fixed." To this point there has been no attempt to, for example, think of a covariate series $\boldsymbol{x}$ as having its own probabilistic structure or joint probabilistic structure with $\boldsymbol{Y}$. To this point (just as in ordinary linear models and regression analysis) we've been operating using only conditional distributions for $\boldsymbol{Y}$ given values of the covariates. Multivariate time series analysis (that would, for example, allow us to treat $\boldsymbol{x}$ as random) is yet to come.

5 Some Considerations in the Practice of Forecasting

When using (the extensions of) ARMA modeling in forecasting we desire (at least)

1. simple models,
2. statistically significant estimates of parameters,
3. good values of "fit" criteria,
4. residuals that look like white noise, and
5. good measures of prediction accuracy.

In the first place, as in all of statistical modeling, one is look for simple, parsimonious, interpretable descriptions of data sets and the scenarios that generate them. Simple models provide convenient mental constructs for describing, predicting, and manipulating the phenomena they represent. The more complex a model is, the less "handy" it is. And further, if one goes too far in the direction of complexity looking for good fit to data in hand, the worse it is likely to do in prediction.

Typically, 0 parameters in a model mean that it reduces to some some simpler model (that doesn't include those parameters). So where a parameter estimate isn't "statistically significant"/"detectably different from 0" there is the indication that some model simpler than the one under consideration might be an adequate description of data in hand. Further, poorly determined parameters in fitted models are often associated with unreliable extrapolations beyond the data set in hand. A parameter that is potentially positive, 0, or negative could often produce quite different extrapolations depending upon very small changes from a current point estimate. One wants to have a good handle on the both sign and magnitude of one's model parameters.

Standard measures of fit that take into account both how well a model will reproduce a data set and how complex it is are Akaike's information criterion and Schwarz's Bayesian information criterion. These are computed by `JMP` and other time series programs and are respectively

$$\begin{aligned} AIC &= -2 \cdot \text{Gaussian loglikelihood} + 2 \cdot \text{number of real parameters} \\ SBC &= -2 \cdot \text{Gaussian loglikelihood} + \ln(n) \cdot \text{number of real parameters} \end{aligned}$$

These criteria penalize model complexity, since small values are desirable.

Plots of estimated autocorrelations for residuals, use of the Ljung-Box statistics Q_h^{LB}, and other ideas of Section 3.4 can and should be brought to bear on the question of whether residuals look like white noise. Where they don't, inferences and predictions based on a fitted model are tenuous and there is additionally the possibility that with more work, more pattern in the data might be identified and exploited.

There are many possible measures of prediction accuracy for a fitted time series model. We proceed to identify a couple and consider how they might be used. For the time being, suppose that for a model fit to an entire available data series $\boldsymbol{Y}_n$ (we'll here not bother to display the estimated model parameters) let

$\hat{y}_t$ be the (best linear) predictor (in the fitted model) of y_t based on $\boldsymbol{Y}_{t-1}$. Measures of prediction accuracy across the data set include

$$R^2 = \frac{\sum_{t=1}^{n}(y_t - \bar{y}_n)^2 - \sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y}_n)^2}$$

exactly as in ordinary regression analysis,

$$\widehat{\sigma^2}$$

(though this is far less directly interpretable than it is, for example, in regression), the mean absolute error

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

the mean absolute percentage error

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{|y_t|} \times 100\%$$

and the so-called "symmetric" mean absolute percentage error

$$SMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{.5\left(|y_t| + |\hat{y}_t|\right)} \times 100\%$$

(The latter is not so symmetric as its name would imply. For example, with positive series, over-prediction by a fixed amount is penalized less than under-prediction by the same amount.) Any or all of these criteria can be computed for a given model and compared to values of the same statistics for alternative models.

An important way of using particularly the statistics of prediction accuracy is to make use of them with so-called "hold-out" samples consisting of the final h elements of $\boldsymbol{Y}_n$. That is, rather than fitting a candidate model form to all n observations, one might fit to only $\boldsymbol{Y}_{n-h}$, holding out h observations. USING THIS 2nd VERSION OF THE FITTED MODEL, let $\hat{y}_t^h$ be the (best linear) predictor (in the fitted model) of y_t based on $\boldsymbol{Y}_{t-1}$. One might then compute hold-out sample versions of accuracy criteria, like

$$\begin{aligned} MAE_h &= \frac{1}{h}\sum_{t=n-h+1}^{n}\left|y_t - \hat{y}_t^h\right| \\ MAPE_h &= \frac{1}{h}\sum_{t=n-h+1}^{n}\frac{\left|y_t - \hat{y}_t^h\right|}{|y_t|} \times 100\% \end{aligned}$$

and

$$SMAPE_h = \frac{1}{h}\sum_{t=n-h+1}^{n}\frac{\left|y_t - \hat{y}_t^h\right|}{.5\left(|y_t| + \left|\hat{y}_t^h\right|\right)} \times 100\%$$

across several values of h (like, for example, $4, 6$, and 8). Note then, that for every h under consideration (including $h = 0$), there is an estimated set of parameters computed on the basis of $\boldsymbol{Y}_{n-h}$, a "regular" version of $MAE, MAPE$, and $SMAPE$ computed across $\boldsymbol{Y}_{n-h}$ using those fitted parameters (that one might call an "in-sample" value of the criterion), and the hold-out versions of the criteria $MAE_h, MAPE_h$, and $SMAPE_h$ computed across $y_{n-h+1}, \ldots, y_n$. One doesn't want parameter estimates to change much across h. (If they do, there is indication that the fitting of the model form is highly sensitive to the final few observed values.) One doesn't want in-sample and hold-out versions of a criterion to be wildly different. (If they are, the holdout version is almost always bigger than the in-sample version, and one must worry about how the model form can be expected to do if more data become available at future periods.) And one doesn't want the criteria to be wildly different across h. (Model effectiveness shouldn't depend strongly upon exactly how many observations are available for fitting.) The `SAS FS`$^{\text{TM}}$ software makes analysis with hold-outs very easy to do.

6 Multivariate Time Series

One motivation for considering multivariate time series (even when forecasting for a univariate $\boldsymbol{Y}$ is in view) is the possibility of applying pieces of multivariate time series methodology to the problem of using covariate series $\boldsymbol{x}$ (now modeled as themselves random) in the forecasting. But to provide a general notation and development, suppose now that $\boldsymbol{Y}$ is $\infty \times m$,

$$\boldsymbol{Y} = \begin{pmatrix} \vdots & & & \vdots \\ y_{-11} & y_{-12} & \cdots & y_{-1m} \\ y_{01} & y_{02} & \cdots & y_{0m} \\ y_{11} & y_{12} & \cdots & y_{1m} \\ \vdots & & & \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \boldsymbol{y}_{-1} \\ \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \\ \vdots \end{pmatrix}$$

where

$$y_{ti} = \text{the value of the } i\text{th series at time } t$$

and the t row of $\boldsymbol{Y}$ is

$$\boldsymbol{y}_t = (y_{t1}, y_{t2}, \ldots, y_{tm})$$

Define

$$\boldsymbol{\mu}_t = \mathrm{E}\boldsymbol{y}_t$$

and

$$\gamma_{ij}(t+h, t) = \mathrm{Cov}(y_{t+h,i}, y_{tj})$$

and

$$\begin{aligned} \underset{m\times m}{\boldsymbol{\Gamma}}(t+h, t) &= \left(\gamma_{ij}(t+h, t)\right)_{\substack{i=1,\ldots,m \\ j=1,\ldots,m}} \\ &= \mathrm{E}\left(\boldsymbol{y}_{t+h} - \boldsymbol{\mu}_{t+h}\right)'(\boldsymbol{y}_t - \boldsymbol{\mu}_t) \end{aligned}$$

This is a matrix of covariances between variables i and j at respective times $t+h$ and t. It is not a covariance matrix and it is not necessarily symmetric.

6.1 Multivariate Second Order Stationary Processes

We'll say that $\boldsymbol{Y}$ is **second order or weakly stationary** if neither $\boldsymbol{\mu}_t$ nor $\boldsymbol{\Gamma}(t+h,t)$ depends upon t. Note that when $\boldsymbol{Y}$ is multivariate second order stationary, each series $\{y_{ti}\}$ is univariate second order stationary. In the event of second order stationarity, we may write simply $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}(h)$, the latter with entries $\gamma_{ij}(h)$, the **cross-covariance functions** between series i and j. Notice that unlike autocovariance functions, the cross-covariance functions are not even, that is, in general $\gamma_{ij}(h) \neq \gamma_{ij}(-h)$. What *is* true however, is that

$$\gamma_{ij}(h) = \gamma_{ji}(-h)$$

so that

$$\boldsymbol{\Gamma}(h) = \boldsymbol{\Gamma}'(-h)$$

We'll call

$$\rho_{ij}(h) = \frac{\gamma_{ij}(h)}{\sqrt{\gamma_{ii}(0)\,\gamma_{jj}(0)}}$$

the **cross-correlation function** between series i and series j. Assembling these in matrices, we write

$$\boldsymbol{R}(h) = \left(\rho_{ij}(h)\right)_{\substack{i=1,\ldots,m \\ j=1,\ldots,m}}$$

and note that (of course) following from the properties of cross-covariances

$$\rho_{ij}(h) = \rho_{ji}(-h)$$

so that

$$\boldsymbol{R}(h) = \boldsymbol{R}'(-h)$$

In a context where what is under consideration are random univariate series $\boldsymbol{Y}$ and $\boldsymbol{x}$ and the intent is to use the covariate $\boldsymbol{x}$ to help forecast $\boldsymbol{Y}$, it will presumably be those cases where

$$\left|\rho_{yx}(h)\right|$$

is large for *positive* h where the $\boldsymbol{x}$ series is of most help in doing the forecasting (one hopes for x that is strongly related to y and is a *leading* indicator of y).

There is a notion of multivariate white noise that is basic to defining tractable models for multivariate time series. That is this. An $\infty \times m$ second order stationary series

$$\boldsymbol{\epsilon} = \begin{pmatrix} \vdots \\ \boldsymbol{\epsilon}_{-1} \\ \boldsymbol{\epsilon}_0 \\ \boldsymbol{\epsilon}_1 \\ \vdots \end{pmatrix}$$

is called white noise with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$ provided

$$\boldsymbol{\mu} = \mathbf{0} \text{ and } \mathbf{\Gamma}(h) = \begin{cases} \mathbf{\Sigma} & \text{if } h = 0 \\ \mathbf{0} & \text{otherwise} \end{cases}$$

Multivariate white noise series can be used to define multivariate series with more complicated dependence structures. To begin, for $\boldsymbol{\epsilon}$ multivariate white noise, if

$$\boldsymbol{y}_t' = \sum_{j=-\infty}^{\infty} \boldsymbol{C}_j \boldsymbol{\epsilon}_{t-j}'$$

where the $m \times m$ matrices $\boldsymbol{C}_j$ have absolutley summable (across j) elements, then the multivariate series

$$\boldsymbol{Y} = \begin{pmatrix} \vdots \\ \boldsymbol{y}_{-1} \\ \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \\ \vdots \end{pmatrix}$$

is called a **linear process**. Where all $\boldsymbol{C}_j$ are $\mathbf{0}$ for $j < 0$, $\boldsymbol{Y}$ is termed a multivariate $\mathbf{MA}(\infty)$ **process**.

There are **mutlivariate ARMA processes** that will be considered below. Any causal multivariate ARMA(p, q) process $\boldsymbol{Y}$ can be represented in "$\mathbf{AR}(\infty)$" form as

$$\boldsymbol{y}_t' - \sum_{j=1}^{\infty} \boldsymbol{A}_j \boldsymbol{y}_{t-j}' = \boldsymbol{\epsilon}_t' \quad \forall t$$

for white noise $\boldsymbol{\epsilon}$ where the matrices $\boldsymbol{A}_j$ have absolutely summable (across j) elements.

6.2 Estimation of Multivariate Means and Correlations for Second Order Stationary Processes

The vector sample mean through period n,

$$\bar{\boldsymbol{y}}_n = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{y}_t$$

is the obvious estimator of the mean $\boldsymbol{\mu}$ of a second order stationary process. Proposition 7.3.1 of BDM provides simple statements of the (mean square error) consistency of that estimator. That is, if $\boldsymbol{Y}$ is second order stationary with mean $\boldsymbol{\mu}$ and covariance function $\mathbf{\Gamma}(h)$, the condition $\gamma_{jj}(n) \to 0 \; \forall j$ is sufficient to guarantee that

$$\mathrm{E}\left(\bar{\boldsymbol{y}}_n - \boldsymbol{\mu}\right)\left(\bar{\boldsymbol{y}}_n - \boldsymbol{\mu}\right)' = \mathrm{E} \sum_{j=1}^{m} \left(\bar{y}_{nj} - \mu_j\right)^2 \to 0$$

And the stronger condition $\sum_{h=-\infty}^{\infty}\left|\gamma_{jj}(h)\right|<\infty\ \forall j$ is sufficient to guarantee the stronger result that

$$n\mathrm{E}\left(\bar{\boldsymbol{y}}_n-\boldsymbol{\mu}\right)\left(\bar{\boldsymbol{y}}_n-\boldsymbol{\mu}\right)'=n\mathrm{E}\sum_{j=1}^{m}\left(\bar{y}_{nj}-\mu_j\right)^2\to\sum_{j=1}^{m}\sum_{h=-\infty}^{\infty}\gamma_{jj}(h)$$

An "obvious" estimator of the matrix cross-covariance function is

$$\hat{\boldsymbol{\Gamma}}_n(h)=\begin{cases}\frac{1}{n}\sum_{t=1}^{n-h}\left(\boldsymbol{y}_{t+h}-\bar{\boldsymbol{y}}_n\right)'\left(\boldsymbol{y}_t-\bar{\boldsymbol{y}}_n\right) & \text{for } 0\le h\le n-1\\ \hat{\boldsymbol{\Gamma}}_n'(-h) & \text{for } -n+1\le h<0\end{cases}$$

which has entries

$$\hat{\gamma}_{ij}(h)=\frac{1}{n}\sum_{t=1}^{n-h}\left(y_{t+h,i}-\bar{y}_{n,i}\right)'\left(y_{t,j}-\bar{y}_{n,j}\right)\quad 0\le h\le n-1$$

From these, estimated cross-correlation functions are

$$\hat{\rho}_{ij}(h)=\frac{\hat{\gamma}_{ij}(h)}{\sqrt{\hat{\gamma}_{ii}(0)\,\hat{\gamma}_{jj}(0)}}$$

and of course the $i=j$ version of this is the earlier autocorrelation function for the ith series.

In looking for predictors x and lags at which those predictors might be useful in forecasting y, one needs some basis upon which to decide when $\left|\hat{\rho}_{yx}(h)\right|$ is of a size that is clearly bigger than would be seen "by chance" if the predictor is unrelated to y. This requires some distributional theory for the sample cross-correlation. Theorem 7.3.1 BDM provides one kind of insight into how big sample cross-correlations can be "by chance" (and what will allow them to be big without indicating that in fact $\left|\rho_{yx}(h)\right|$ is big). That result is as follows.

Suppose that for iid $\epsilon_{t,1}$ with mean 0 and variance σ_1^2 independent of iid $\epsilon_{t,2}$ with mean 0 and variance σ_2^2,

$$y_t=\sum_{k=-\infty}^{\infty}\alpha_k\epsilon_{t-k,1}\quad\forall t$$

and

$$x_t=\sum_{k=-\infty}^{\infty}\beta_k\epsilon_{t-k,2}\quad\forall t$$

for $\sum_{k=-\infty}^{\infty}|\alpha_k|<\infty$ and $\sum_{k=-\infty}^{\infty}|\beta_k|<\infty$ (so that y_t and x_t are independent and thus uncorrelated linear processes). Then for large n and $h\neq k$

$$\sqrt{n}\begin{pmatrix}\hat{\rho}_{yx}(h)\\ \hat{\rho}_{yx}(k)\end{pmatrix}\dot{\sim}\mathrm{MVN}_2\left(\mathbf{0},\begin{pmatrix}v & c\\ c & v\end{pmatrix}\right)$$

for

$$v = \sum_{j=-\infty}^{\infty} \rho_{yy}(j)\,\rho_{xx}(j) \quad \text{and} \quad c = \sum_{j=-\infty}^{\infty} \rho_{yy}(j)\,\rho_{xx}(j+k-h)$$

Of course, an alternative statement of the limiting marginal distribution for $\hat{\rho}_{yx}(h)$ for any h when the series are independent, is that $\hat{\rho}_{yx}(h)$ is approximately normal with mean 0 and standard deviation

$$\sqrt{\frac{1}{n}\left(1 + 2\sum_{j=1}^{\infty} \rho_{yy}(j)\,\rho_{xx}(j)\right)}$$

This result indicates that even when two series are independent and model cross-correlations are thus 0, they can have large *sample* cross-correlations if their autocorrelation functions more or less "line up nicely." There is thus in general no simple cut-off value that separates big from small sample cross-correlations. But notice that if (at least) one of the series y_t or x_t is white noise, $v = 1$ and we have the result that for any h,

$$\sqrt{n}\hat{\rho}_{yx}(h) \dot{\sim} \mathrm{N}(0,1)$$

So with high probability

$$\left|\hat{\rho}_{yx}(h)\right| < \frac{2}{\sqrt{n}}$$

and in this case there *is* a simple yardstick (that doesn't vary with the character of the non-white-noise process) against which one can judge the statistical significance of sample cross-correlations.

The practical usefulness of this insight in data analysis is this. If one fits a time series model to one of the series $\boldsymbol{Y}$ or $\boldsymbol{x}$ and then computes residuals, sample autocorrelations between the other series and the residual series should be small (typically less than $2/\sqrt{n}$ in magnitude) if the original $\boldsymbol{Y}$ and $\boldsymbol{x}$ series are independent second order stationary series. The notion of reducing one of the series to residuals before examining sample cross-correlations is known as **whitening** or **pre-whitening**.

To be absolutely explicit about what is being suggested here, suppose that some ARIMA (or SARIMA) model has been fit to $\boldsymbol{Y}$, i.e. one has found parameters so that for some difference operator $\mathcal{D}^*$,

$$\mathcal{D}^*\ \boldsymbol{Y} \sim \mathrm{ARMA}(p,q) \quad ,$$

$\hat{e}_t^*$ for $t = 1, 2, \ldots, n$ is the corresponding set of (standardized) residuals, and $\mathcal{D}^*\boldsymbol{X}$ is the corresponding differenced covariate series. Then (provided the $\mathcal{D}^*\boldsymbol{X}$ series is itself stationary) looking at values of

$$\hat{\rho}_{\hat{e}^*,\mathcal{D}^*x}(h)$$

is a way of looking for lags at which the predictor might be an effective covariate for forecasting of y_t. If, for example, $\hat{\rho}_{\hat{e}^*,\mathcal{D}^*x}(4)$ greatly exceeds $2/\sqrt{n}$ in

magnitude, using a differenced version of the predictor lagged by 4 periods in a form for the mean of $\mathcal{D}^* \boldsymbol{Y}$ (that is, βx^*_{t-4} in the mean of y_t) is a promising direction in a transfer function model search.

6.3 Multivariate ARMA Processes

These seem far less useful in practice than their univariate counterparts, but if for no other reason to see how the univariate ideas generalize, we will briefly consider them.

6.3.1 Generalities

We will say that a mean $\mathbf{0}$ second order stationary multivariate process model for $\boldsymbol{Y}$ is ARMA(p,q) provided there exist $m \times m$ matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$ and a covariance matrix $\boldsymbol{\Sigma}$ such that for every t

$$\boldsymbol{y}'_t - \boldsymbol{\Phi}_1 \boldsymbol{y}'_{t-1} - \cdots - \boldsymbol{\Phi}_p \boldsymbol{y}'_{t-p} = \boldsymbol{\epsilon}'_t + \boldsymbol{\Theta}_1 \boldsymbol{\epsilon}'_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{\epsilon}'_{t-q}$$

for

$$\boldsymbol{\epsilon} = \begin{pmatrix} \vdots \\ \boldsymbol{\epsilon}_{-1} \\ \boldsymbol{\epsilon}_0 \\ \boldsymbol{\epsilon}_1 \\ \vdots \end{pmatrix}$$

white noise with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{Y}$ is mean $\boldsymbol{\mu}$ ARMA(p,q) provided

$$\boldsymbol{Y} - \begin{pmatrix} \vdots \\ \boldsymbol{\mu} \\ \boldsymbol{\mu} \\ \boldsymbol{\mu} \\ \vdots \end{pmatrix}$$

is mean $\mathbf{0}$ ARMA(p,q).

A standard simple example of a multivariate ARMA process is the AR(1) case. This is the case where

$$\boldsymbol{y}'_t = \boldsymbol{\Phi} \boldsymbol{y}'_{t-1} + \boldsymbol{\epsilon}'_t \quad \forall t$$

for an $m \times m$ matrix $\boldsymbol{\Phi}_1$ and white noise $\boldsymbol{\epsilon}$. As it turns out, provided all eigenvalues of $\boldsymbol{\Phi}$ are less than 1 in magnitude, one may (in exact analogy with the $m = 1$ case) write

$$\boldsymbol{y}'_t = \sum_{j=0}^{\infty} \boldsymbol{\Phi}^j \boldsymbol{\epsilon}'_{t-j} \quad \forall t$$

The components of $\boldsymbol{\Phi}^j$ are absolutely summable and $\boldsymbol{Y}$ is a linear process with an MA(∞) form.

Causality and invertibility for multivariate ARMA processes are exactly analogous to those properties for univariate ARMA processes. First, a multivariate ARMA process is **causal** if there exist $m \times m$ matrices $\boldsymbol{\Psi}_j$ with absolutely summable components such that

$$\boldsymbol{y}_t' = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\epsilon}_{t-j}' \quad \forall t$$

When there is causality, exactly as indicated in display (35) for univariate cases, the matrices $\boldsymbol{\Psi}_j$ can be computed from the recursion

$$\begin{aligned} \boldsymbol{\Psi}_j &= \boldsymbol{\Theta}_j + \sum_{k=1}^{p} \boldsymbol{\Phi}_k \boldsymbol{\Psi}_{j-k} \quad \text{for } j = 0, 1, \ldots \\ \text{where } \boldsymbol{\Theta}_0 &= \boldsymbol{I}, \boldsymbol{\Theta}_j = \boldsymbol{0} \quad \text{for } j > q \quad \text{and } \boldsymbol{\Psi}_j = 0 \quad \text{for } j < 0 \end{aligned}$$

Clearly, provided all eigenvalues of $\boldsymbol{\Phi}$ are less than 1 in magnitude, a multivariate AR(1) process is causal with $\boldsymbol{\Psi}_j = \boldsymbol{\Phi}^j$.

Then, a multivariate ARMA process is **invertible** if there exist $m \times m$ matrices $\boldsymbol{\Pi}_j$ with absolutely summable components such that

$$\boldsymbol{\epsilon}_t' = \sum_{j=0}^{\infty} \boldsymbol{\Pi}_j \boldsymbol{y}_{t-j}' \quad \forall t$$

When there is invertibility, exactly as indicated display (36) for univariate cases, the matrices $\boldsymbol{\Pi}_j$ can be computed from the recursion

$$\begin{aligned} \boldsymbol{\Pi}_j &= -\boldsymbol{\Phi}_j - \sum_{k=1}^{q} \boldsymbol{\Theta}_k \boldsymbol{\Pi}_{j-k} \quad \text{for } j = 0, 1, \ldots \\ \text{where } \boldsymbol{\Phi}_0 &= -\boldsymbol{I}, \boldsymbol{\Phi}_j = \boldsymbol{0} \quad \text{for } j > p \quad \text{and } \boldsymbol{\Pi}_j = 0 \quad \text{for } j < 0 \end{aligned}$$

An annoying feature of the multivariate ARMA development is that even restriction to causal and invertible representations of multivariate ARMA processes is not sufficient to deal with lack-of-identifiability issues. That is, a single mean and covariance structure for a multivariate time series $\boldsymbol{Y}$ can come from more than one causal and invertible ARMA model. To see this is true, consider an example (essentially one on page 243 of BDM) of an $m = 2$ case of AR(1) structure with

$$\boldsymbol{\Phi} = \begin{bmatrix} 0 & c \\ 0 & 0 \end{bmatrix}$$

for $|c| < 1$. Since for $j \geq 2$ it is the case that $\boldsymbol{\Phi}^j = \boldsymbol{0}$,

$$\boldsymbol{y}_t' = \boldsymbol{\epsilon}_t' + \boldsymbol{\Phi} \boldsymbol{\epsilon}_{t-1}' \quad \forall t$$

and $\boldsymbol{Y}$ also has an MA(1) representation. One way out of this lack-of-identifiability problem that is sometimes adopted is to consider only pure AR multivariate processes, i.e. only ARMA$(p, 0)$ models.

6.3.2 Covariance Functions and Prediction

The matrix function of lagged cross covariances $\mathbf{\Gamma}(h) = \mathrm{E}\left(\boldsymbol{y}_{t+h} - \boldsymbol{\mu}_{t+h}\right)'(\boldsymbol{y}_t - \boldsymbol{\mu}_t)$ for a causal multivariate ARMA process turns out to be (in direct analogy to the univariate form for the autocovariance function of univariate linear processes (10))

$$\mathbf{\Gamma}(h) = \sum_{j=0}^{\infty} \mathbf{\Psi}_{h+j} \mathbf{\Sigma} \mathbf{\Psi}_j'$$

and often this series converges fast enough to use a truncated version of it for practical computation of $\mathbf{\Gamma}(h)$. Alternatively (in exact analogy to what is in Section 3.2 for univariate ARMA processes) one can use a recursion to compute values of $\mathbf{\Gamma}(h)$. That is,

$$\mathbf{\Gamma}(j) - \sum_{k=1}^{p} \mathbf{\Phi}_k \mathbf{\Gamma}(j-k) = \sum_{k=j}^{q} \mathbf{\Theta}_k \mathbf{\Sigma} \mathbf{\Psi}_{k-j} \quad \forall j \geq 0$$

Using the fact that $\mathbf{\Gamma}(-h) = \mathbf{\Gamma}'(h)$ the instances of this equation for $j = 0, 1, \ldots, p$ can be solved simultaneously for $\mathbf{\Gamma}(0), \mathbf{\Gamma}(1), \ldots, \mathbf{\Gamma}(p)$. Then the recursion can then be used to find $\mathbf{\Gamma}(h)$ for $h > p$. In any case, there is a computational path from the ARMA model parameters $\mathbf{\Phi}_j, \mathbf{\Theta}_j$, and $\mathbf{\Sigma}$ to the function $\mathbf{\Gamma}(h)$.

Then, at least conceptually, the path to best linear prediction in multivariate ARMA models is clear. After properly arranging the elements of $\boldsymbol{Y}$ that are to be predicted and those to be used for prediction into a single column vector and writing the corresponding covariance matrix (in terms of values of $\mathbf{\Gamma}(h)$) the usual formulas for a multivariate normal conditional mean and conditional covariance matrix lead to best linear predictors and their prediction covariance matrices. (All else beyond this is detail of interest to time series specialists and those who will program this in special cases and need to make use of computational shortcuts available in particular models and prediction problems.)

To make concrete what is intended in the previous paragraph, consider the prediction of $\boldsymbol{y}_{n+1}$ from $\boldsymbol{Y}_n$ in any multivariate second order stationary process with mean $\boldsymbol{\mu}$ and matrix function of lagged cross covariances $\mathbf{\Gamma}(h)$ (including multivariate ARMA processes). If we rearrange the $(n+1) \times m$ matrix $\boldsymbol{Y}_{n+1}$ into an $((n+1)m)$-dimensional column vector as

$$\boldsymbol{Y}_{n+1}^{*} = \begin{pmatrix} \boldsymbol{y}_1' \\ \boldsymbol{y}_2' \\ \vdots \\ \boldsymbol{y}_n' \\ \boldsymbol{y}_{n+1}' \end{pmatrix}$$

second order stationarity produces

$$\mathrm{E}\boldsymbol{Y}_{n+1}^{*} = \begin{pmatrix} \boldsymbol{\mu}' \\ \boldsymbol{\mu}' \\ \vdots \\ \boldsymbol{\mu}' \\ \boldsymbol{\mu}' \end{pmatrix}$$

and

$$\mathrm{Var}\left(\boldsymbol{Y}_{n+1}^{*}\right) = \begin{pmatrix} \boldsymbol{\Gamma}(0) & \boldsymbol{\Gamma}(-1) & \boldsymbol{\Gamma}(-2) & \cdots & \boldsymbol{\Gamma}(-n) \\ \boldsymbol{\Gamma}(1) & \boldsymbol{\Gamma}(0) & \boldsymbol{\Gamma}(-1) & \cdots & \boldsymbol{\Gamma}(-n+1) \\ \boldsymbol{\Gamma}(2) & \boldsymbol{\Gamma}(1) & \boldsymbol{\Gamma}(0) & \cdots & \boldsymbol{\Gamma}(-n+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}(n) & \boldsymbol{\Gamma}(n-1) & \boldsymbol{\Gamma}(n-2) & \cdots & \boldsymbol{\Gamma}(0) \end{pmatrix}$$

(Remember that $\boldsymbol{\Gamma}(h) = \boldsymbol{\Gamma}'(-h)$ so that this matrix really is symmetric.) But now under a Gaussian assumption, it's in theory perfectly possible to identify the mean and variance of the conditional distribution of the last m entries of $\boldsymbol{Y}_{n+1}^{*}$ given the first nm of them. The conditional mean is the best linear predictor of $\boldsymbol{y}_{n+1}'$ based on $\boldsymbol{Y}_n$, and the conditional covariance matrix is the prediction covariance matrix.

6.3.3 Fitting and Forecasting with Multivariate AR(p) Models

Presumably because there is ambiguity of representation of second order structures using multivariate ARMA models unless one further narrows the set of possibilities, BDM deal specifically with the class of multivariate AR(p) models in their Section 7.6. There are basically two lines of discussion in that section. In the first place, narrowing one's focus to pure AR models makes it possible to identify efficient ways to compute a $\mathbf{0}$ mean Gaussian likelihood

$$f\left(\boldsymbol{Y}_n | \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p, \boldsymbol{\Sigma}\right)$$

thereby potentially making Gaussian-based inference practical in such models. The authors note that one must select an order (p) based on data, and suggest (AR model) selection criteria like

$$AICC = -2\ln f\left(\boldsymbol{Y}_n | \widehat{\boldsymbol{\Phi}_1}, \ldots, \widehat{\boldsymbol{\Phi}_p}, \widehat{\boldsymbol{\Sigma}}\right) + \frac{2\left(pm^2+1\right)nm}{nm - pm^2 - 2}$$

to make sure that one takes account of the very large model complexity implicit in anything by a very low order multivariate ARMA model.

The second line of discussion concerns prediction in multivariate AR(p) processes. For $n \geq p$, s-step-ahead multivariate AR(p) forecasts are very simple. The $\mathbf{0}$ mean version is this. First (with $\hat{\boldsymbol{y}}_{n+s}$ the s-step-ahead from time n)

$$\hat{\boldsymbol{y}}_{n+1}' = \boldsymbol{\Phi}_1 \boldsymbol{y}_n' + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{n+1-p}'$$

and then

$$\begin{aligned}\hat{\boldsymbol{y}}'_{n+2} &= \boldsymbol{\Phi}_1\hat{\boldsymbol{y}}'_{n+1} + \boldsymbol{\Phi}_2\boldsymbol{y}'_n + \cdots + \boldsymbol{\Phi}_p\boldsymbol{y}'_{n-p} \text{ and} \\ \hat{\boldsymbol{y}}'_{n+3} &= \boldsymbol{\Phi}_1\hat{\boldsymbol{y}}'_{n+2} + \boldsymbol{\Phi}_2\hat{\boldsymbol{y}}'_{n+1} + \boldsymbol{\Phi}_3\boldsymbol{y}'_n + \cdots + \boldsymbol{\Phi}_p\boldsymbol{y}'_{n-p-1}\end{aligned}$$

etc. Prediction covariance matrices for s-step-ahead multivariate AR(p) forecasts are simply

$$\sum_{j=0}^{s-1} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}'_j$$

6.3.4 Multivariate ARIMA (and SARIMA) Modeling and Co-Integration

There is the possibility of using a multivariate ARMA model after differencing a set of m series making up a multivariate $\boldsymbol{Y}$. That is, for $\mathcal{D}^*$ some difference operator and

$$\boldsymbol{Y} = \left(\boldsymbol{Y}^1, \boldsymbol{Y}^2, \ldots, \boldsymbol{Y}^m\right)$$

where each $\boldsymbol{Y}^j$ is an $\infty \times 1$ series, we can agree to write $\mathcal{D}^*\boldsymbol{Y}$ for the $\infty \times m$ multivariate series $\left(\mathcal{D}^*\boldsymbol{Y}^1, \mathcal{D}^*\boldsymbol{Y}^2, \ldots, \mathcal{D}^*\boldsymbol{Y}^m\right)$ and adopt a multivariate ARMA model for $\mathcal{D}^*\boldsymbol{Y}$. (Depending upon the exact nature of $\mathcal{D}^*$) $\mathcal{D}^*\boldsymbol{Y}$ is then multivariate ARIMA (or SARIMA).

There is another idea for producing a stationary series from $\boldsymbol{Y}$, that (instead of differencing down columns) focuses on combining the columns of $\boldsymbol{Y}$ to produce a second order stationary series. This is the concept of **co-integration**. The motivation is that the columns of $\boldsymbol{Y}$ may not be second order stationary, while some linear combination of the elements of $\boldsymbol{y}_t$ moves in a consistent/stationary way that can be modeled using the methods discussed thus far.

The version of the idea discussed by BDM is this. Say that $\infty \times m$ multivariate series $\boldsymbol{Y}$ is **integrated of order** d if $\mathcal{D}^d\boldsymbol{Y}$ is second order stationary and $\mathcal{D}^{d-1}\boldsymbol{Y}$ is not. Then if $\boldsymbol{Y}$ is integrated of order d and there exists an $m \times 1$ vector $\boldsymbol{\alpha}$ such that the $\infty \times 1$ series $\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{\alpha}$ is integrated of order less than d, we'll say that $\boldsymbol{Y}$ is **co-integrated with co-integration vector** $\boldsymbol{\alpha}$.

A fairly concrete example of co-integration provided on BDM page 255 is this. Suppose that $\infty \times 1$ series $\boldsymbol{Y}$ satisfies

$$\mathcal{D}\boldsymbol{Y} = \boldsymbol{\epsilon}$$

for $\boldsymbol{\epsilon}$ mean 0 white noise. $\boldsymbol{Y}$ is a random walk and is not second order stationary.

Let $\boldsymbol{Z}$ be univariate white noise independent of $\boldsymbol{Y}$ and define

$$\boldsymbol{W} = \boldsymbol{Y} + \boldsymbol{Z}$$

Then the multivariate series $(\boldsymbol{Y}, \boldsymbol{W})$ is integrated of order 1. $(\mathcal{D}(\boldsymbol{Y}, \boldsymbol{W}) = (\boldsymbol{\epsilon}, \boldsymbol{\epsilon} + \mathcal{D}\boldsymbol{Z})$ and it's fairly easy to show that this is second order stationary.) But consider $\boldsymbol{\alpha} = (-1, 1)'$ and

$$(\boldsymbol{Y}, \boldsymbol{W})\boldsymbol{\alpha} = -\boldsymbol{Y} + (\boldsymbol{Y} + \boldsymbol{Z}) = \boldsymbol{Z}$$

which is second order stationary. So $(\boldsymbol{Y}, \boldsymbol{W})$ is co-integrated with co-integration vector $(-1, 1)'$. The basic idea is that the component series $\boldsymbol{Y}$ and $\boldsymbol{W}$ are non-stationary, but their difference is stationary. They both wander about like random walks, but they "wander together" and their difference is in some sense stable.

7 Heuristic Time Series Decompositions/Analyses and Forecasting Methods

There are any number of heuristics for time series analysis and forecasting that have no firm basis in probability modeling, but have nevertheless proved useful over years of practice. We consider a few of these, concentrating on ones discussed in Section 1.5 and Chapter 9 of BDM.

7.1 "Classical" Decomposition of $\boldsymbol{Y}_n$

A conceptually attractive decomposition of an observed univariate time series is as

$$y_t = m_t + s_t + n_t \quad \text{for } t = 1, 2, \ldots, n \tag{42}$$

where for some $d > 0$

$$s_{t+d} = s_t \ \forall t \quad \text{and} \quad \sum_{j=1}^{d} s_j = 0 \tag{43}$$

Here m_t represents "trend," s_t is a "seasonal component," and what's left over, n_t, is "noise." More or less standard/classical fitting of this structure proceeds as follows.

First, we compute a preliminary version of trend using a moving average matched to d and designed so that under the conceptualization (42) and (43) it is unaffected by the seasonal components. That is, for d even, say $d = 2q$, define

$$\tilde{m}_t = \frac{1}{d}\left(\frac{1}{2}\left(y_{t-q} + y_{t+q}\right) + \sum_{j=-q+1}^{q-1} y_{t+j}\right)$$

and for d odd, say $d = 2q + 1$, define

$$\tilde{m}_t = \frac{1}{d}\sum_{j=-q}^{q} y_{t+j}$$

(Under these definitions, each $\tilde{m}_t$ includes exactly one copy of each s_j for $j = 1, 2, \ldots, d$ as a summand, and since $\sum_{j=1}^{d} s_j = 0$ values of these do not affect the values $\tilde{m}_t$.)

Second, we compute preliminary versions of fitted seasonal components as

$$\tilde{s}_j = \text{average of } \{y_j - \tilde{m}_j, y_{j+d} - \tilde{m}_{j+d}, y_{j+2d} - \tilde{m}_{j+2d}, \ldots\} \quad \text{for } j = 1, 2, \ldots, d$$

Now these won't necessarily add to 0, so define final versions of fitted seasonal components as

$$\hat{s}_j = \tilde{s}_j - \frac{1}{d}\sum_{j=1}^{d} \tilde{s}_j \quad \text{for } j = 1, 2, \ldots, d$$

Then deseasonalized/seasonally adjusted values of the data are

$$d_t = y_t - \hat{s}_t$$

Next, we compute final versions of the fitted trend, say $\hat{m}_t$, using some smoothing method applied to the deseasonalized values, d_t. Possible modern methods of smoothing that could be used include smoothing splines and kernel smoothers. Classical methods of smoothing traditionally applied include moving average smoothing of the form

$$\hat{m}_t = \frac{1}{2q+1}\sum_{j=-q}^{q} d_{t+j} \quad ,$$

polynomial regressions, and exponential smoothing, where

$$\hat{m}_t = \alpha d_t + (1-\alpha)\,\hat{m}_{t-1}$$

for some fixed $\alpha \in (0, 1)$.

Then, of course, $\hat{n}_t$ is what is left over

$$\hat{n}_t = y_t - \hat{m}_t - \hat{s}_t$$

7.2 Holt-Winters Smoothing/Forecasting

There are both ordinary and seasonal versions of this methodology. We'll describe both.

7.2.1 No Seasonality

The basic idea here is a kind of adaptive/exponentially smoothed linear extrapolation/forecasting where

$$\hat{y}_{n+s} = \hat{a}_n + \hat{b}_n s \tag{44}$$

for $\hat{a}_n$ and $\hat{b}_n$ respectively fitted/smoothed "level" and "slope" of the time series at time $t = n$. Beginning at time $t = 2$, one might define

$$\hat{a}_2 = y_2 \quad \text{and} \quad \hat{b}_2 = y_2 - y_1$$

Then for $3 \leq t \leq n$ one defines fitted values recursively by

$$\hat{y}_t = \hat{a}_{t-1} + \hat{b}_{t-1}\,(1)$$

(this is the fitted level at time $t-1$ incremented by 1 times the fitted slope at time $t-1$). Levels and slopes are updated recursively in an "exponential smoothing" way, i.e. via

$$\hat{a}_t = \alpha y_t + (1-\alpha)\,\hat{y}_t$$

for some fixed $\alpha \in (0,1)$ and

$$\hat{b}_t = \beta\,(\hat{a}_t - \hat{a}_{t-1}) + (1-\beta)\,\hat{b}_{t-1}$$

for some fixed $\beta \in (0,1)$.

The smoothing parameters α and β control how fast the fitted level and slope can change. They might be chosen to minimize a criterion like

$$Q(\alpha,\beta) = \sum_{t=3}^{n} (y_t - \hat{y}_t)^2$$

BDM claim that this version of H-W forecasting (that ultimately makes use of the linear extrapolation (44)) is for large n essentially equivalent to forecasting using an ARIMA model

$$\mathcal{D}^2 \boldsymbol{Y} = \left(\mathcal{I} - (2-\alpha-\alpha\beta)\,\mathcal{B} + (1-\alpha)\,\mathcal{B}^2\right)\boldsymbol{\epsilon}$$

7.2.2 With Seasonality

A seasonal version of the Holt-Winters algorithm produces extrapolations/forecasts

$$\hat{y}_{n+s} = \hat{a}_n + \hat{b}_n s + \hat{c}_{n+s} \tag{45}$$

where $\hat{a}_n$ and $\hat{b}_n$ are respectively fitted/smoothed "level" and "slope" of the time series at time $t=n$ and $\hat{c}_{n+s} = \hat{c}_{n+s-kd}$ for k the smallest non-negative integer for which $n+s-kd \leq n$ and $\hat{c}_{n-d+1}, \hat{c}_{n-d+2}, \ldots, \hat{c}_{n-1}, \hat{c}_n$ are fitted/smoothed versions of "seasonal components" of the series relevant at time $t=n$. Beginning at time $t = d+1$, we define

$$\hat{a}_{d+1} = y_{d+1} \quad \text{and} \quad \hat{b}_{d+1} = (y_{d+1} - y_1)/d$$

and take

$$\hat{c}_t = y_t - \left(y_1 + \hat{b}_{d+1}\,(t-1)\right) \quad \text{for} \quad t = 2, \ldots, d+1$$

Then for $t > d+1$,fitted values are updated recursively via

$$\hat{y}_t = \hat{a}_{t-1} + \hat{b}_{t-1}\,(1) + \hat{c}_{t-d}$$

(this is the fitted level at time $t-1$ incremented by 1 times the fitted slope at time $t-1$ plus fitted seasonal component for time t). Levels, slopes and seasonal components are updated recursively in an "exponential smoothing" way, i.e. via

$$\hat{a}_t = \alpha\,(y_t - \hat{c}_{t-d}) + (1-\alpha)\left(\hat{a}_{t-1} + \hat{b}_{t-1}\,(1)\right)$$

for some fixed $\alpha \in (0,1)$ and

$$\hat{b}_t = \beta\left(\hat{a}_t - \hat{a}_{t-1}\right) + (1-\beta)\,\hat{b}_{t-1}$$

for some fixed $\beta \in (0,1)$ and

$$\hat{c}_t = \gamma\left(y_t - \hat{a}_t\right) + (1-\gamma)\,\hat{c}_{t-d}$$

for some fixed $\gamma \in (0,1)$.

As in the nonseasonal case, parameters α, β, and γ control how fast the fitted level, slope, and seasonal components can change. They might be chosen to minimize a criterion like

$$Q(\alpha,\beta,\gamma) = \sum_{t=d+2}^{n} \left(y_t - \hat{y}_t\right)^2$$

Further, much as for the nonseasonal case of Holt-Winters forecasting, BDM suggest that the present seasonal version of H-W forecasting (that ultimately makes use of the extrapolation (45)) is for large n essentially equivalent to forecasting using an ARIMA model specified by

$$\mathcal{D}\mathcal{D}_d \boldsymbol{Y} = \left(\sum_{j=0}^{d-1} \mathcal{B}^j + \gamma(1-\alpha)\,\mathcal{B}^d\mathcal{D} - (2-\alpha-\alpha\beta)\sum_{j=1}^{d} \mathcal{B}^j + (1-\alpha)\sum_{j=2}^{d+1} \mathcal{B}^j\right)\boldsymbol{\epsilon}$$

8 Direct Modeling of the Autocovariance Function

From some points of view, the "Box-Jenkins/ARMA" enterprise is somewhat dissatisfying. One is essentially mapping sets of parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ^2 to autocovariance functions, trying to find sets of parameters that reproduce an observed/estimated form. But the link between the parameters and the character of autocovariance functions is less than completely transparent. (For one thing, even the process variance $\gamma(0)$ fails to be related to the parameters in a simple fashion!) That at least suggests the possibility of taking a more direct approach to modeling the autocovariance function. There is nothing about the ARMA structure that makes it necessary for (or even particularly suited to) the additional differencing and regression elements of standard times series and forecasting methods. If one could identify a form for and estimate parameters of $\gamma(s)$ directly, all of the differencing, regression, forecasting, and model checking material discussed thus far would remain in force. (Of course, the real impediment to application here is the need for appropriate software to implement more direct analysis of time series dependence structures.)

Here we outline what could at least in theory be done in general, and is surely practically feasible "from scratch" in small problems if an analyst is at all effective at statistical computing (e.g. in `R`). It begins with the construct of

an autocorrelation function for a stochastic process in continuous (rather than discrete) time.

Basically *any* symmetric non-negative definite real-valued function of a single real variable, say $f(t)$, can serve as an autocovariance function for a stochastic process of a continuous variable t. When such a function is divided by its value at 0 (forming $\phi(t) = f(t)/f(0)$) a valid autocorrelation function for a stochastic process of the continuous variable t is formed. People in spatial statistics have identified a number of convenient forms ϕ (and further noted that one rich source of such functions consists of real-valued characteristic functions associated with distributions on $\Re$ that are symmetric about 0). So (coming from a variety of sources) Table 1 provides an example set of basic autocorrelation functions for stochastic process of a real variable, t. Any one of these function ϕ may be rescaled (in time) by replacing its argument t with ct for a positive constant (parameter) c. And we then note that when evaluated at only integer values s, these functions of real t serve as autocorrelation functions for time series.

Next we observe that the product of two or more autocorrelation functions is again an autocorrelation function and that (weighted) averages of two or more autocorrelation functions are again autocorrelation functions. Finally, we can be reminded that we know the simple forms for white noise and AR(1) autocorrelation functions. All this ultimately means that we have at our disposal a wide variety of basic autocorrelation function forms $\phi(cs)$ using entries of Table 1 plus the (1-parameter) AR(1) form (that can take 2 basic shapes, exponentially declining and possibly oscillating) and the white noise form that can be multiplied or averaged together (introducing weights and time scalings, c, as parameters) to produce new parametric families of autocorrelation functions. It then seems entirely possible to build a form matching the general shape of a typical sample autocorrelation function $\hat{\rho}_n(s)$ at small to moderate lags. Then all parameters of the correlation function (any c's, and any AR(1) lag 1 correlation, and any weights) plus (an interpretable) variance $\gamma(0)$ can become parameters of a covariance matrix for $\boldsymbol{Y}_n$ to be estimated by (for example) Gaussian maximum likelihood.

Table 1: Some Basic Autocorrelation Functions

$\phi(t)$	Origin/Source
$\cos \lvert t \rvert$	Chf of distribution with mass $\frac{1}{2}$ at each of ± 1
$\frac{\sin \lvert t \rvert}{\lvert t \rvert}$	Chf of U$(-1, 1)$ distribution
$\frac{\sin^2 \lvert t \rvert}{t^2}$	Chf of triangular distribution on $(-2, 2)$
e^{-t^2}	Chf of N$(0, 2)$ distribution
$e^{-\lvert t \rvert}$	Chf standard Cauchy distribution
$e^{-\lvert t \rvert^\nu}$ (for a $0 < \nu \leq 2$)	
$\frac{1}{1+t^2}$	Chf of symmetrized Exp(1) distribution
$\frac{1}{(1+t^2)^\beta}$ (for a $\beta > 0$)	spatial statistics
$(1 + \lvert t \rvert) e^{-\lvert t \rvert}$	spatial statistics
$\left(1 + \lvert t \rvert + \frac{t^2}{3}\right) e^{-\lvert t \rvert}$	spatial statistics
$(1 - \lvert t \rvert)_+$	
$(1 - \lvert t \rvert)_+^3 (3 \lvert t \rvert + 1)$	

9 Spectral Analysis of Second Order Stationary Time Series

This material is about decomposing second order stationary series and their correlation/covariance structures into periodic components. As usual, we suppose throughout that $\boldsymbol{Y}$ is second order stationary with autocovariance function $\gamma(s)$.

9.1 Spectral Distributions

To begin, suppose that $\gamma(s)$ is absolutely summable, that is

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty \tag{46}$$

In this case the (apparently complex-valued) function of the real variable λ,

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma(h) \tag{47}$$

is well-defined and is called the "**spectral density**" for $\boldsymbol{Y}$. In fact (remembering that for θ real, $e^{i\theta} = \cos\theta + i\sin\theta$ and that $\gamma(-h) = \gamma(h)$) this function is real-valued, and

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) \cos(\lambda h) \tag{48}$$

The fact that $\cos\theta$ is an even function implies that $f(-\lambda) = f(\lambda)$. Since $\cos\theta$ has period 2π, $f(\lambda)$ is periodic with period $2\pi/k$ for an integer $k \geq 1$, and it suffices to consider $f(\lambda)$ as defined on $[-\pi, \pi]$. Further, there is a technical argument in BDM that shows that $f(\lambda) \geq 0$, to some degree justifying the "density" terminology.

What is most interesting is that the autocovariances can be recovered from the spectral density. To see this, write

$$\begin{aligned}
\int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda)\, d\lambda &= \int_{-\pi}^{\pi} e^{ik\lambda} \left(\frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma(h) \right) d\lambda \\
&= \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) \int_{-\pi}^{\pi} e^{i(k-h)\lambda} d\lambda \\
&= \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) \int_{-\pi}^{\pi} \cos((k-h)\lambda)\, d\lambda \\
&= \gamma(k)
\end{aligned}$$

Now to this point all of this has been developed under the assumption (46) that the autocovariance function is absolutely summable. We can define spectral densities for some cases where this restriction is not met, by simply saying

that if there is an even real function $f(\lambda) \geq 0$ for $-\pi \leq \lambda \leq \pi$ with

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)\, d\lambda = \int_{-\pi}^{\pi} \cos(h\lambda)\, f(\lambda)\, d\lambda \tag{49}$$

then we'll call $f(\lambda)$ the spectral density for the process.

A reasonable question is: "What functions can serve as spectral densities?" Proposition 4.1.1 of BDM says that a non-negative real-valued function $f(\lambda)$ on $[-\pi, \pi]$ is the spectral density for a second order stationary process if and only if $f(-\lambda) = f(\lambda)$ and

$$\int_{-\pi}^{\pi} f(\lambda)\, d\lambda < \infty$$

Notice that this latter means that

$$g(\lambda) = \frac{f(\lambda)}{\int_{-\pi}^{\pi} f(\lambda)\, d\lambda}$$

is the pdf of a continuos distribution on $[-\pi, \pi]$ that is symmetric about 0.

Then observe that when $f(\lambda)$ is a spectral density, the relationship (49) can be thought of in terms involving expected values. That is, with

$$\sigma^2 = \gamma(0) = \mathrm{Var} y_t = \int_{-\pi}^{\pi} f(\lambda)\, d\lambda$$

and $g(\lambda) = f(\lambda)/\sigma^2$ the pdf on $[-\pi, \pi]$ derived from $f(\lambda)$, for $W \sim g$, the relationship (49) is

$$\gamma(h) = \sigma^2 \mathrm{E} e^{ihW} = \sigma^2 \mathrm{E} \cos(hW) \tag{50}$$

(So, incidentally, $\rho(h) = \mathrm{E} e^{ihW} = \mathrm{E} \cos(hW)$ and the lag h autocorrelation is the g mean of $\cos(hW)$.)

Now, not all autocovariance functions for second order stationary processes have spectral densities. But it is always possible to provide a representation of the form (50). That is, a version of the BDM Theorem 4.1.1 says that $\gamma(h)$ is an autocovariance function for a second order stationary process if and only if there is a symmetric probability distribution G on $[-\pi, \pi]$ such that for $W \sim G$ the relationship (50) holds. The (generalized, since its total mass is σ^2, that is potentially different from 1) distribution $\sigma^2 G$ is called the **spectral distribution** for the process. Where G is continuous, the spectral distribution has the spectral density g. But it's also perfectly possible for G to be discrete (or neither continuous nor discrete).

The spectral distribution of a process not only provides a theoretical tool for reconstructing the autocovariance function, it is typically interpreted as giving insight into how "rough" realizations of a time series are likely to be, and how those realizations might be thought of as made from periodic (sinusoidal) components. Rationale for this thinking can be developed through a series of examples.

Consider first the case of "random sinusoids." That is, consider a second order stationary process defined by

$$y_t = \sum_{j=1}^{k} \left(A_j \cos\left(\omega_j t\right) + B_j \sin\left(\omega_j t\right) \right) \tag{51}$$

for some set of frequencies $0 < \omega_1 < \omega_2 < \cdots < \omega_k < \pi$, and uncorrelated mean 0 random variables $A_1, \ldots, A_k, B_1, \ldots, B_k$ where $\text{Var}A_j = \text{Var}B_j = \sigma_j^2$. It's a trigonometric fact that

$$A_j \cos\left(\omega_j t\right) + B_j \sin\left(\omega_j t\right) = \sqrt{A_j^2 + B_j^2} \sin\left(\omega_j t + \phi_j\right)$$

for $\phi_j \in (-\pi, \pi]$ the unique angle satisfying $\sin\left(\phi_j\right) = A_j / \sqrt{A_j^2 + B_j^2}$ and $\cos\left(\phi_j\right) = B_j / \sqrt{A_j^2 + B_j^2}$. So the form (51) is a sum of sine functions with random weights $\sqrt{A_j^2 + B_j^2}$, fixed frequencies ω_j, and random phase shifts/offsets ϕ_j. In general, a large σ_j^2 will produce a large amplitude $\sqrt{A_j^2 + B_j^2}$ or weight on the sinusoid of frequency ω_j.

Then, as it turns out, the process (51) has autocovariance function

$$\gamma\left(h\right) = \sum_{j=1}^{k} \sigma_j^2 \cos\left(\omega_j h\right)$$

so that $\sigma^2 = \gamma\left(0\right) = \sum_{j=1}^{k} \sigma_j^2$. But notice that for a discrete random variable W with

$$P\left[W = \omega_j\right] = P\left[W = -\omega_j\right] = \frac{1}{2}\left(\frac{\sigma_j^2}{\sigma^2}\right)$$

it is the case that

$$\text{E}e^{ihW} = \text{E}\cos\left(hW\right) = \sum_{j=1}^{k} \left(\frac{\sigma_j^2}{\sigma^2}\right) \cos\left(\omega_j h\right)$$

so that $\gamma\left(h\right) = \sigma^2 \text{E}e^{ihW}$ and the generalized distribution that is σ^2 times the probability distribution of W is the (discrete) spectral distribution of the process. This spectral distribution places large mass on those frequencies for which realizations of the process will tend to have corresponding large sinusoidal components. In particular, in cases where large frequencies have large associated masses, one can expect realizations of the process to be "rough" and have fast local variation.

As a second example of a spectral distribution, consider the spectral density

$$f\left(\lambda\right) = \frac{\sigma^2}{2\pi} \quad \text{for } \lambda \in [-\pi, \pi]$$

This prescribes a "flat" spectral distribution/a flat spectrum. Here, for $h \neq 0$

$$\gamma(h) = \frac{\sigma^2}{2\pi} \int_{-\pi}^{\pi} \cos(h\lambda)\, d\lambda = 0$$

and we see that this is the white noise spectral density. (Analogy to physics, where white light is electromagnetic radiation that has constant intensities of components at all wavelengths/frequencies, thus provides motivation for the "white noise" name.) This spectral distribution produces extremely rough realizations for $\boldsymbol{Y}$, spreading appreciable mass across large frequencies.

AR(1) and MA(1) processes have fairly simple and illuminating spectral densities. In the first case (where in this section the white noise variance will be denoted by η^2 thereby reserving the notation σ^2 for the variance of y_t) for $\lambda \in [-\pi, \pi]$

$$\begin{aligned} f(\lambda) &= \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \phi^{|h|} \left(\frac{\eta^2}{1-\phi^2} \right) \cos(\lambda h) \\ &= \frac{\eta^2}{2\pi \left(1 - 2\phi\cos(\lambda) + \phi^2\right)} \end{aligned}$$

It's easy to verify that for ϕ near 1, this density puts small mass at large frequencies and AR(1) realizations are relatively smooth, while for ϕ near -1 the density puts large mass at large frequencies and AR(1) realizations are relatively rough, involving "fast random" oscillations.

The MA(1) spectral density is for $\lambda \in [-\pi, \pi]$

$$\begin{aligned} f(\lambda) &= \frac{1}{2\pi} \sum_{h=-1}^{1} \gamma(h) \cos(\lambda h) \\ &= \frac{\eta^2}{2\pi} \left(1 + 2\theta\cos(\lambda) + \theta^2\right) \end{aligned}$$

It is easy to verify that for θ near 1, this density puts small mass at large frequencies and MA(1) realizations are relatively smooth, while for θ near -1 the density puts large mass at large frequencies and MA(1) realizations are relatively rough, again involving "fast random" oscillations.

9.2 Linear Filters and Spectral Densities

Suppose that $\mathcal{L} = \sum_{t=-\infty}^{\infty} \psi_t \mathcal{B}^t$ is a time-invariant linear filter with absolutely summable coefficients (i.e. with $\sum_{t=-\infty}^{\infty} |\psi_t| < \infty$). Proposition 1 on page 16 says how the autocovariance function for $\mathcal{L}\boldsymbol{Y}$ is related to that of $\boldsymbol{Y}$. In the event that $\boldsymbol{Y}$ has a spectral density, it is possible to also provide a very simple formula for the spectral density of $\mathcal{L}\boldsymbol{Y}$. A small amount of new notation must be prepared in order to present this.

Corresponding to $\mathcal{L}$ define the (complex-valued) function of $\lambda \in [-\pi, \pi]$

$$T_{\mathcal{L}}(\lambda) = \sum_{t=-\infty}^{\infty} \psi_t e^{-it\lambda} = \sum_{t=-\infty}^{\infty} \psi_t (\cos(-t\lambda) + i\sin(-t\lambda)) = \sum_{t=-\infty}^{\infty} \psi_t (\cos(t\lambda) - i\sin(t\lambda))$$

This is sometimes called the "**transfer function**" of the linear filter $\mathcal{L}$. Related to this is the (real non-negative) so-called **power transfer function** of the linear filter

$$|T_{\mathcal{L}}(\lambda)|^2 = \left(\sum_{t=-\infty}^{\infty} \psi_t \cos(t\lambda)\right)^2 + \left(\sum_{t=-\infty}^{\infty} \psi_t \sin(t\lambda)\right)^2$$

With this notation, it is possible to show that spectral densities are related by

$$f_{\mathcal{L}\boldsymbol{Y}}(\lambda) = |T_{\mathcal{L}}(\lambda)|^2 f_{\boldsymbol{Y}}(\lambda) \tag{52}$$

The relationship (52) has several interesting immediate consequences. For example, consider the seasonal (lag s) difference operator, $\mathcal{D}_s$. This has only two non-zero coefficients, $\psi_0 = 1$ and $\psi_s = -1$. So the corresponding transfer function is

$$T_{\mathcal{D}_s}(\lambda) = 1 - e^{-is\lambda}$$

Then, for integer k,

$$T_{\mathcal{D}_s}\left(k\left(\frac{2\pi}{s}\right)\right) = 0$$

so that for integer k

$$f_{\mathcal{D}_s\boldsymbol{Y}}\left(k\left(\frac{2\pi}{s}\right)\right) = \left|T_{\mathcal{D}_s}\left(k\left(\frac{2\pi}{s}\right)\right)\right|^2 f_{\boldsymbol{Y}}\left(k\left(\frac{2\pi}{s}\right)\right) = 0$$

One might interpret this to mean that $\mathcal{D}_s\boldsymbol{Y}$ has no sinusoidal components with periods that are divisors of s. The seasonal differencing in some sense removes from the distribution of $\boldsymbol{Y}$ periodic patterns that complete some number of full cycles in exactly s time periods. This is is completely consistent with the standard data-analytic motivation of seasonal differencing.

As a second important application of relationship (52) consider finding the spectral density for a general ARMA process. Suppose that

$$\Phi(\mathcal{B})\boldsymbol{Y} = \Theta(\mathcal{B})\boldsymbol{\epsilon}$$

Then the spectral density for $\Phi(\mathcal{B})\boldsymbol{Y}$ is

$$f_{\Phi(\mathcal{B})\boldsymbol{Y}}(\lambda) = \left|T_{\Phi(\mathcal{B})}(\lambda)\right|^2 f_{\boldsymbol{Y}}(\lambda)$$

and the spectral density for $\Theta(\mathcal{B})\boldsymbol{\epsilon}$ is

$$\begin{aligned} f_{\Theta(\mathcal{B})\boldsymbol{\epsilon}}(\lambda) &= \left|T_{\Theta(\mathcal{B})}(\lambda)\right|^2 f_{\boldsymbol{\epsilon}}(\lambda) \\ &= \left|T_{\Theta(\mathcal{B})}(\lambda)\right|^2 \left(\frac{\sigma^2}{2\pi}\right) \end{aligned}$$

Equating these two spectral densities and solving for $f_{\boldsymbol{Y}}(\lambda)$ we then have

$$f_{\boldsymbol{Y}}(\lambda) = \frac{\sigma^2}{2\pi}\left(\frac{\left|T_{\Theta(\mathcal{B})}(\lambda)\right|^2}{\left|T_{\Phi(\mathcal{B})}(\lambda)\right|^2}\right)$$

9.3 Estimating a Spectral Density

If one assumes that $\boldsymbol{Y}$ has a spectral density, a natural question is how one might estimate it based on $\boldsymbol{Y}_n$. The basic tool typically used in such estimation is the so-called "**periodogram**" of $\boldsymbol{Y}_n$. This is the function defined on $[-\pi, \pi]$ by

$$\begin{aligned} I_n(\lambda) &= \frac{1}{n}\left|\sum_{t=1}^{n} y_t e^{-it\lambda}\right|^2 \\ &= \frac{1}{n}\left|\sum_{t=1}^{n} y_t\left(\cos(t\lambda) - i\sin(t\lambda)\right)\right|^2 \end{aligned}$$

The periodogram turns out to be a first empirical approximation of $2\pi f(\lambda)$. This can be motivated by Proposition 4.2.1 of BDM. This result says that for any $\omega \in (-\pi, \pi]$ of the form $\omega = 2\pi k/n$ for k a non-zero integer (a so-called Fourier frequency),

$$I_n(\omega) = \sum_{|h|<n} \hat{\gamma}_n(h) e^{-ih\omega} = \sum_{|h|<n} \hat{\gamma}_n(h)\cos(h\omega) \tag{53}$$

Recalling the opening definitions of a spectral density (47) and (48), the relationship (53) for Fourier frequencies then suggests that $I_n(\lambda)$ might in general approximate $2\pi f(\lambda)$. But as it turns out, it is necessary to modify the periodogram by smoothing in order to produce a consistent estimator of the function $2\pi f(\lambda)$.

For $m(n)$ an integer (that typically grows with n), a non-negative symmetric weight function $w_n(j)$ on the integers (that can depend upon n) with

$$\sum_{j=-m(n)}^{m(n)} w_n(j) = 1$$

and function of $\lambda \in (-\pi, \pi]$

$$g(n, \lambda) = \text{the multiple of } 2\pi/n \text{ closest to } \lambda$$

an estimator of $f(\lambda)$ based on a smoothed periodogram is

$$\tilde{f}(\lambda) = \frac{1}{2\pi}\sum_{|j|\le m(n)} w_n(j) I_n\left(g(n,\lambda) + \frac{2\pi j}{n}\right)$$

(where relationship (53) could be used to find the values of I_n at Fourier frequencies that are needed here). In practice, the form of weights $w_n(j)$ used is chosen to smooth I_n, but to not smooth it "too much."

Various choices of neighborhood size $m(n)$ and weights $w_n(j)$ lead to consistency results for the estimator $\tilde{f}(\lambda)$. One example is the choice

$$m(n) = \text{the greatest integer in } \sqrt{n}$$

and

$$w_j(n) = \frac{1}{2m(n)+1}$$

For this choice, $\tilde{f}(\lambda)$ is essentially $1/2\pi$ times an arithmetic average of I_n evaluated at the roughly $2\sqrt{n}+1$ Fourier frequencies closest to λ.

10 State Space Models

So-called state space formulations of time series models and corresponding Kalman recursions provide a flexible and effective general methodology for modeling and analysis. They provide 1) a unified treatment of many standard models, 2) recursive prediction, filtering, and smoothing, 3) recursive likelihood calculations, 4) natural handling of missing values, and 5) direct generalization to non-Gaussian and non-linear models. BDM provides a description of these in its Chapter 8, and what follows here is a combination of a retelling of the BDM development and some class notes of Ken Ryan whose origin is probably in an ISU course of Jay Breidt.

10.1 Basic State Space Representations

Here we are going to abandon/modify some of the multivariate time series notation we used in Section 6. It doesn't seem particularly helpful in the present context to make much use of operator notation, $\Re^\infty$ vectors, or $\infty \times m$ representations of m-variate time series. We *will* want to be able to index multivariate observations by time and will most naturally prefer to write them for fixed time t as column vectors (rather than as row vectors as we did before). So here, unless specifically indicated to the contrary, vectors are column vectors. For example, $\boldsymbol{y}_t$ will be a column vector of observations at time t (in contrast to the convention we used earlier that would make it a row vector).

The basic sate space formulation operates on two interrelated sets of recursions, the first for a system "state" and the second for a corresponding "observation." (Most simply, one conceives of a stochastic evolution of the state and a clouded/noisy perception of it.) We'll write the state (or transition) equation/recursion as

$$\underset{v\times 1}{\boldsymbol{x}_{t+1}} = \underset{v\times v}{\boldsymbol{F}_t}\ \underset{v\times 1}{\boldsymbol{x}_t} + \underset{v\times 1}{\boldsymbol{v}_t} \tag{54}$$

and the observation (or measurement) equation/recursion as

$$\underset{w\times 1}{\boldsymbol{y}_t} = \underset{w\times v}{\boldsymbol{G}_t}\ \underset{v\times 1}{\boldsymbol{x}_t} + \underset{w\times 1}{\boldsymbol{w}_t} \tag{55}$$

for (at least for the time being, non-random) matrices $\boldsymbol{F}_t$ and $\boldsymbol{G}_t$, and mean $\boldsymbol{0}$ random vectors $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$. We will assume that the error vectors

$$\begin{pmatrix} \boldsymbol{v}_t \\ \boldsymbol{w}_t \end{pmatrix}$$

are uncorrelated with each other and (where only time $t > 0$ is considered) with the initial state, $\boldsymbol{x}_1$. The fixed t covariance matrix for the errors will be written as

$$\mathrm{E}\begin{pmatrix} \boldsymbol{v}_t \\ \boldsymbol{w}_t \end{pmatrix}(\boldsymbol{v}_t', \boldsymbol{w}_t') = \begin{pmatrix} \underset{v\times v}{\boldsymbol{Q}_t} & \underset{v\times w}{\boldsymbol{S}_t} \\ \underset{w\times v}{\boldsymbol{S}_t'} & \underset{w\times w}{\boldsymbol{R}_t} \end{pmatrix}$$

and $\boldsymbol{F}_t, \boldsymbol{G}_t, \boldsymbol{Q}_t, \boldsymbol{R}_t$ and $\boldsymbol{S}_t$ are system matrices. In the event they don't change with t, the system is "time-invariant." This structure covers a wide variety of both second order stationary and other models for the observable $\boldsymbol{y}_t$.

A simple (perhaps the archetypal) example of this formalism is the "random walk plus noise." Let

$$x_{t+1} = x_t + v_t$$

for $\{v_t\}$ a univariate mean 0 variance σ_v^2 white noise sequence and

$$y_t = x_t + w_t$$

for $\{w_t\}$ a univariate mean 0 variance σ_w^2 white noise sequence uncorrelated with the v sequence. This model is of state space form (54) and (55) with $F_t = 1, G_t = 1, Q_t = \sigma_v^2, R_t = \sigma_w^2$, and $S_t = 0$.

A slight generalization of the example of cointegration in Section 6.3.4 can be formulated as a second example of the state space structure. That is, for $\{v_t\}$ and $\{w_t\}$ uncorrelated univariate mean 0 white noise sequences with respective variances σ_v^2 and σ_w^2, let

$$\begin{aligned} d_{t+1} &= d_t + v_t \quad \text{and} \\ p_t &= \gamma d_t + w_t \end{aligned}$$

let

$$\boldsymbol{y}_t = \begin{pmatrix} d_t \\ p_t \end{pmatrix}$$

Then $\{\boldsymbol{y}_t\}$ is integrated of order 1. To see this, note that while the random walk $\{d_t\}$ is not second order stationary (so that $\{\boldsymbol{y}_t\}$ is not second order stationary),

$$\boldsymbol{y}_t - \boldsymbol{y}_{t-1} = \begin{pmatrix} d_t - d_{t-1} \\ p_t - p_{t-1} \end{pmatrix} = \begin{pmatrix} v_{t-1} \\ \gamma\left(d_t - d_{t-1}\right) + w_t - w_{t-1} \end{pmatrix} = \begin{pmatrix} v_{t-1} \\ \gamma v_{t-1} + w_t - w_{t-1} \end{pmatrix}$$

has entries that *are* second order stationary.

In state space form we can write

$$\boldsymbol{y}_t = \begin{pmatrix} 1 \\ \gamma \end{pmatrix} d_t + \begin{pmatrix} 0 \\ w_t \end{pmatrix}$$

and with $x_t = d_t$

$$x_{t+1} = 1x_t + v_t$$

This is the case of the state space from (54) and (55) with $F_t = 1, \boldsymbol{G}_t = (1, \gamma)', \boldsymbol{w}_t = (0, w_t), Q_t = \sigma_v^2, \boldsymbol{S}_t = (0, 0)$, and

$$\boldsymbol{R}_t = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_w^2 \end{pmatrix}$$

And to complete the "cointegration" story for this example, note that $(-\gamma, 1)$ is a cointegration vector since

$$(-\gamma, 1)\, \boldsymbol{y}_t = -\gamma d_t + p_t = w_t$$

and the w sequence is second order stationary.

As it turns out, causal invertible ARMA models can be realized as the distributions of $\boldsymbol{y}_t$ in state space models. In this preliminary look at the state space formulation, consider the simple AR(2), MA(1), and ARMA(1, 1) instances of this truth. First, consider an AR(2) model specified in difference equation form as

$$x_{t+1} = \phi_1 x_t + \phi_2 x_{t-1} + \epsilon_{t+1}$$

For

$$\boldsymbol{x}_t = \begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix}, \boldsymbol{F}_t = \begin{pmatrix} \phi_1 & \phi_2 \\ 0 & 0 \end{pmatrix}, \text{ and } \boldsymbol{v}_t = \begin{pmatrix} \epsilon_{t+1} \\ 0 \end{pmatrix}$$

state equation (54) and the instance of observation equation (55) with $\boldsymbol{G}_t = (1, 0)$ and w_t with variance 0 produces a state space representation of the AR(2) model for $y_t = x_t$ with

$$\boldsymbol{Q}_t = \begin{pmatrix} \sigma_v^2 & 0 \\ 0 & 0 \end{pmatrix}$$

$\boldsymbol{S}_t = (0, 0)'$, and $R_t = 0$. (Note that this extends in obvious fashion to AR(p) cases using p-dimensional state vectors consisting of p lags of y's.)

To produce a state space representation of an MA(1) model, consider a 2-dimensional state vector

$$\boldsymbol{x}_t = \begin{pmatrix} x_{t,1} \\ x_{t,2} \end{pmatrix}$$

and a state equation

$$\boldsymbol{x}_{t+1} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \boldsymbol{x}_t + \begin{pmatrix} \epsilon_{t+1} \\ \theta\epsilon_{t+1} \end{pmatrix}$$

which is clearly the version of equation (54) with

$$\boldsymbol{F}_t = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \text{ and } \boldsymbol{v}_t = \begin{pmatrix} \epsilon_{t+1} \\ \theta\epsilon_{t+1} \end{pmatrix}$$

Then $\boldsymbol{G}_t = (1,0)$ and w_t with variance 0 produces a state space representation of the MA(1) model for $y_t = (1,0)\,\boldsymbol{x}_t = x_{t,1} + \epsilon_t = x_{t-1,2} + \epsilon_t = \theta\epsilon_{t-1} + \epsilon_t$. (A similar argument could be mounted for the MA(q) case using $(q+1)$-dimensional state vectors.)

Then, to cover the ARMA(1,1) possibility expressed in difference equation form as

$$y_t = \phi y_{t-1} + \theta\epsilon_{t-1} + \epsilon_t$$

suppose that $\{x_t\}$ is a univariate causal AR(1) process satisfying the operator equation

$$(\mathcal{I} - \phi\mathcal{B})\,\boldsymbol{x} = \boldsymbol{\epsilon}$$

for 0 mean white noise $\boldsymbol{\epsilon}$. We may then use

$$\boldsymbol{x}_t = \begin{pmatrix} x_{t-1} \\ x_t \end{pmatrix}, \boldsymbol{F}_t = \begin{pmatrix} 0 & 1 \\ 0 & \phi \end{pmatrix}, \text{ and } \boldsymbol{v}_t = \begin{pmatrix} 0 \\ \epsilon_{t+1} \end{pmatrix}$$

in the state equation (54) and $\boldsymbol{G}_t = (\theta, 1)$ and w_t with variance 0 in the observation equation (55), producing a model for the observable satisfying

$$\begin{aligned}
y_t &= \theta x_{t-1} + x_t \\
&= \theta \sum_{j=0}^{\infty} \phi^j \epsilon_{t-1-j} + \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j} \\
&= \theta\epsilon_{t-1} + \theta \sum_{j=1}^{\infty} \phi^j \epsilon_{t-1-j} + \sum_{j=1}^{\infty} \phi^j \epsilon_{t-j} + \epsilon_t \\
&= \phi \left(\theta \sum_{j=1}^{\infty} \phi^{j-1} \epsilon_{t-1-j} + \sum_{j=1}^{\infty} \phi^{j-1} \epsilon_{t-j} \right) + \theta\epsilon_{t-1} + \epsilon_t \\
&= \phi \left(\theta \sum_{j=0}^{\infty} \phi^j \epsilon_{t-2-j} + \sum_{j=0}^{\infty} \phi^j \epsilon_{t-1-j} \right) + \theta\epsilon_{t-1} + \epsilon_t \\
&= \phi y_{t-1} + \theta\epsilon_{t-1} + \epsilon_t
\end{aligned}$$

10.2 "Structural" Models

It is possible to use state space models to produce solid probabilistic formulations of the heuristic classical decompositions and Holt-Winters thinking discussed in Section 7. These are often known as "structural models" for time series and are discussed here.

To begin, consider the conceptual decomposition (42)

$$y_t = m_t + s_t + n_t$$

where m_t represents an approximate "local level," s_t represents a seasonal effect (any consecutive d of which sum to roughly 0), and n_t represents small "noise." For simplicity of exposition, let's here consider the case of quarterly data, i.e.

the case of $d = 4$. A way of letting both the local level and the form of seasonal effects evolve across time is to employ a state space model related to this decomposition. To this end, let

$$\boldsymbol{x}_t = \begin{pmatrix} m_t \\ s_t \\ s_{t-1} \\ s_{t-2} \end{pmatrix}, \ \boldsymbol{F}_t = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \text{ and } \ \boldsymbol{v}_t = \begin{pmatrix} v_t^m \\ v_t^s \\ 0 \\ 0 \end{pmatrix}$$

for (mean 0 variance σ_m^2 and σ_s^2) uncorrelated white noise sequences $\{v_t^m\}$ and $\{v_t^s\}$ and consider the model with state equation (54) and observation equation

$$y_t = (1, 1, 0, 0)\, \boldsymbol{x}_t + n_t$$

for a (mean 0 variance σ_n^2) white noise sequence $\{n_t\}$ uncorrelated with the state equation errors. Then with $\boldsymbol{G}_t = (1, 1, 0, 0)$ and $w_t = n_t$ this is of state space form (55) and the natural covariance matrix for $(\boldsymbol{v}_t', w_t)'$ is $\boldsymbol{diag}\left(\sigma_m^2, \sigma_s^2, 0, 0, \sigma_n^2\right)$.

A generalization of this development is one where the effect of a "local slope" is added to the local level producing the representation

$$y_t = m_t + b_t \cdot 1 + s_t + n_t$$

(This thinking is much like that leading to seasonal Holt-Winters forecasting.) Let

$$\boldsymbol{x}_t = \begin{pmatrix} m_t \\ b_t \\ s_t \\ s_{t-1} \\ s_{t-2} \end{pmatrix}, \ \boldsymbol{F}_t = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \text{ and } \ \boldsymbol{v}_t = \begin{pmatrix} v_t^m \\ v_t^b \\ v_t^s \\ 0 \\ 0 \end{pmatrix}$$

for uncorrelated (mean 0 variance $\sigma_m^2, \sigma_b^2, \sigma_s^2$) white noise sequences $\left\{v_t^m\right\}, \left\{v_t^b\right\}$, and $\{v_t^s\}$ and consider the model with state equation (54) and observation equation

$$y_t = (1, 1, 1, 0, 0)\, \boldsymbol{x}_t + n_t$$

for a (mean 0 variance σ_n^2) white noise sequence $\{n_t\}$ uncorrelated with the state equation errors. Then with $\boldsymbol{G}_t = (1, 1, 1, 0, 0)$ and $w_t = n_t$ this is of state space form (55) and the natural covariance matrix for $(\boldsymbol{v}_t', w_t)'$ is $\boldsymbol{diag}\left(\sigma_m^2, \sigma_b^2, \sigma_s^2, 0, 0, \sigma_n^2\right)$.

10.3 The Kalman Recursions

The computational basis of application of state space models is a set of recursions for conditional means and variances (that ultimately come from assuming that all $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are jointly Gaussian). Assume that for all t, $\boldsymbol{S}_t = \boldsymbol{0}$. In what follows, we'll write $\boldsymbol{x}_{t|n}$ for the conditional mean of $\boldsymbol{x}_t$ given all $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ and $\hat{\boldsymbol{x}}_{t+1}$ for $\boldsymbol{x}_{t+1|t}$.

The **start-up** of the Kalman computations requires a (prior) distribution for $\boldsymbol{x}_1$. Let $\hat{\boldsymbol{x}}_1$ be the mean of that distribution and $\boldsymbol{\Omega}_1$ be the covariance matrix for that distribution. Beginning with those values, for $t = 1, 2, \ldots, n$ there are then

- **Innovation Recursions**

$$\begin{aligned} \boldsymbol{I}_t &= \boldsymbol{y}_t - \boldsymbol{G}_t \hat{\boldsymbol{x}}_t \quad \text{and} \\ \boldsymbol{\Delta}_t &= \boldsymbol{G}_t \boldsymbol{\Omega}_t \boldsymbol{G}_t' + \boldsymbol{R}_t \end{aligned}$$

  (for the innovations and their covariance matrices),

- **Update/Filter Recursions**

$$\begin{aligned} \boldsymbol{x}_{t|t} &= \hat{\boldsymbol{x}}_t + \boldsymbol{\Omega}_t \boldsymbol{G}_t' \boldsymbol{\Delta}_t^- \boldsymbol{I}_t \quad \text{and} \\ \boldsymbol{\Omega}_{t|t} &= \boldsymbol{\Omega}_t - \boldsymbol{\Omega}_t \boldsymbol{G}_t' \boldsymbol{\Delta}_t^- \boldsymbol{G}_t \boldsymbol{\Omega}_t \end{aligned}$$

  (where $\boldsymbol{\Delta}_t^-$ is any generalized inverse of $\boldsymbol{\Delta}_t$, the recursions giving the conditional means of states and their error covariance matrices $\boldsymbol{\Omega}_{t|t} = \mathrm{E}\big(\boldsymbol{x}_t - \boldsymbol{x}_{t|t}\big)\big(\boldsymbol{x}_t - \boldsymbol{x}_{t|t}\big)'$), and

- **Prediction Recursions**

$$\begin{aligned} \hat{\boldsymbol{x}}_{t+1} &= \boldsymbol{F}_t \boldsymbol{x}_{t|t} \quad \text{and} \\ \boldsymbol{\Omega}_{t+1} &= \boldsymbol{F}_t \boldsymbol{\Omega}_{t|t} \boldsymbol{F}_t' + \boldsymbol{Q}_t \end{aligned}$$

  (for one-step-ahead predictions of states and their error covariance matrices $\boldsymbol{\Omega}_t = \mathrm{E}(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t)(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t)'$).

It is possible to cycle through these recursions in the order above for a given t and produce innovations, updates, and predictions (and their associated covariance matrices) for all of $t = 1, 2, \ldots, n$.

As a completely unrealistic but correspondingly simple illustration of the Kalman calculations, consider the trivial case of a state space model with state equation

$$x_{t+1} = x_t (= \mu)$$

and observation equation

$$y_t = x_t + w_t$$

for $\{w_t\}$ a mean 0 variance σ^2 white noise sequence. Here $F_t = 1, G_t = 1, Q_t = 0$, and $R_t = \sigma^2$ (and $S_t = 0$). As start-up assumptions, suppose that we employ a prior mean of $\hat{x}_1 = \mu_0$ and a prior variance of $\Omega_1 = \sigma_0^2$. Then the $t = 1$ innovation and variance are

$$\begin{aligned} I_1 &= y_1 - 1 \cdot \hat{x}_1 = y_1 - \mu_0 \quad \text{and} \\ \Delta_1 &= 1 \cdot \sigma_0^2 \cdot 1 + \sigma^2 = \sigma_0^2 + \sigma^2 \end{aligned}$$

Next, the $t = 1$ Kalman filter update and corresponding error variance are

$$\begin{aligned} x_{1|1} &= \mu_0 + \sigma_0^2 \cdot 1 \left(\frac{1}{\sigma_0^2 + \sigma^2} \right) (y_1 - \mu_0) = \frac{\sigma^2 \mu_0 + \sigma_0^2 y_1}{\sigma_0^2 + \sigma^2} \quad \text{and} \\ \Omega_{1|1} &= \sigma_0^2 - \sigma_0^2 \cdot 1 \cdot \left(\frac{1}{\sigma_0^2 + \sigma^2} \right) \cdot 1 \cdot \sigma_0^2 = \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2} \end{aligned}$$

And the $t = 1$ prediction and prediction variance are

$$\begin{aligned} \hat{x}_2 &= 1 \cdot x_{1|1} = \frac{\sigma^2 \mu_0 + \sigma_0^2 y_1}{\sigma_0^2 + \sigma^2} \quad \text{and} \\ \Omega_2 &= 1 \cdot \Omega_{1|1} \cdot 1 + 0 = \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2} \end{aligned}$$

Of course, with $t = 1$ recursions completed, the $t = 2$ cycle can begin with innovation and variance

$$\begin{aligned} I_2 &= y_2 - 1 \cdot x_{1|1} = y_2 - \frac{\sigma^2 \mu_0 + \sigma_0^2 y_1}{\sigma_0^2 + \sigma^2} \quad \text{and} \\ \Delta_2 &= 1 \cdot \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2} \cdot 1 + \sigma^2 = \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2} + \sigma^2 \end{aligned}$$

and so on.

10.4 Implications and Extensions of the Kalman Recursions

There are important direct consequences of the basic Kalman recursions just presented.

10.4.1 Likelihood-Based Inference

Under an assumption of joint normality for all $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, (and continuing to assume that all $\boldsymbol{S}_t = \boldsymbol{0}$) the natural log of the joint pdf of the observables (the $\boldsymbol{y}_t$'s) is (for $\boldsymbol{\Lambda}$ a vector of parameters in the matrices $\boldsymbol{F}_t$, $\boldsymbol{G}_t$, $\boldsymbol{Q}_t$, and $\boldsymbol{R}_t$) of the form

$$\ln f(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n | \boldsymbol{\Lambda}) = -\frac{nw}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^{n} \ln \det \boldsymbol{\Delta}_t - \frac{1}{2} \sum_{t=1}^{n} \boldsymbol{I}_t' \boldsymbol{\Delta}_t^{-} \boldsymbol{I}_t$$

For fixed $\boldsymbol{\Lambda}$ this depends only on the innovations and the corresponding variances that can be computed from the Kalman recursions. But this Gaussian log-likelihood (function of $\boldsymbol{\Lambda}$) then translates directly to the possibility of maximum likelihood estimation of $\boldsymbol{\Lambda}$, and an estimated covariance matrix corresponding to the estimates based on the negative Hessian of this function evaluated at the MLE. (Of course, all these will typically need to be determined numerically.)

10.4.2 Filtering and Prediction

After fitting a state space model, one can use it to make predictions and (prediction limits based on them and) prediction covariance matrices. Both $\boldsymbol{x}$'s and $\boldsymbol{y}$'s might be predicted. Consider first prediction of $\boldsymbol{x}$'s.

Prediction for $\boldsymbol{x}_n$ is known as "filtering." It is covered directly by the Kalman filtering recursions.

One-step-ahead prediction of $\boldsymbol{x}_{n+1}$ is based directly on the Kalman filtering and the state equation as

$$\hat{\boldsymbol{x}}_{n+1} = \boldsymbol{F}_n \boldsymbol{x}_{n|n}$$

Two-step prediction is based on

$$\boldsymbol{x}_{n+2|n} = \boldsymbol{F}_{n+1}\hat{\boldsymbol{x}}_{n+1} = \boldsymbol{F}_{n+1}\boldsymbol{F}_n \boldsymbol{x}_{n|n}$$

And in general, s-step prediction is based on

$$\boldsymbol{x}_{n+s|n} = \boldsymbol{F}_{n+s-1}\boldsymbol{F}_{n+s-2}\cdots\boldsymbol{F}_n \boldsymbol{x}_{n|n}$$

Prediction variances for these predictors can be obtained recursively. With

$$\boldsymbol{\Omega}_n^{(s)} = \mathrm{E}\left(\boldsymbol{x}_{n+s} - \boldsymbol{x}_{n+s|n}\right)\left(\boldsymbol{x}_{n+s} - \boldsymbol{x}_{n+s|n}\right)'$$

and using the convention $\boldsymbol{\Omega}_n^{(1)} = \boldsymbol{\Omega}_{n+1}$, for $s \geq 2$ it is the case that

$$\boldsymbol{\Omega}_n^{(s)} = \boldsymbol{F}_{n+s-1}\boldsymbol{\Omega}_n^{(s-1)}\boldsymbol{F}'_{n+s-1} + \boldsymbol{Q}_{n+s-1}$$

Consider then prediction of $\boldsymbol{y}$'s. s-step prediction of $\boldsymbol{y}_{n+s}$ is based on

$$\boldsymbol{y}_{n+s|n} = \boldsymbol{G}_{n+s}\boldsymbol{x}_{n+s|n}$$

This has corresponding error covariance matrix

$$\boldsymbol{\Delta}_n^{(s)} = \mathrm{E}\left(\boldsymbol{y}_{n+s} - \boldsymbol{y}_{n+s|n}\right)\left(\boldsymbol{y}_{n+s} - \boldsymbol{y}_{n+s|n}\right)'$$

satisfying

$$\boldsymbol{\Delta}_n^{(s)} = \boldsymbol{G}_{n+s}\boldsymbol{\Omega}_n^{(s)}\boldsymbol{G}'_{n+s} + \boldsymbol{R}_{n+s}$$

10.4.3 Smoothing

This is prediction of $\boldsymbol{x}_t$ from the observations $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ for $n > t$, producing both

$$\boldsymbol{x}_{t|n} \quad \text{and} \quad \boldsymbol{\Omega}_{t|n} = \mathrm{E}\left(\boldsymbol{x}_t - \boldsymbol{x}_{t|n}\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_{t|n}\right)'$$

BDM argues that these can be computed recursively, beginning with

$$\boldsymbol{x}_{t|t-1} = \hat{\boldsymbol{x}}_t \quad \text{and} \quad \boldsymbol{\Omega}_{t,t} = \boldsymbol{\Omega}_{t|t-1} = \boldsymbol{\Omega}_t$$

(from the Kalman prediction/filtering recursions). Then for $n = t, t+1, t+2, \ldots$

$$\begin{aligned}
\boldsymbol{\Omega}_{t,n+1} &= \boldsymbol{\Omega}_{t,n}\left(\boldsymbol{F}_n - \boldsymbol{F}_n\boldsymbol{\Omega}_n\boldsymbol{G}_n'\boldsymbol{\Delta}_n^-\boldsymbol{G}_n\right)' \\
\boldsymbol{\Omega}_{t|n} &= \boldsymbol{\Omega}_{t|n-1} - \boldsymbol{\Omega}_{t,n}\boldsymbol{G}_n'\boldsymbol{\Delta}_n^-\boldsymbol{G}_n\boldsymbol{\Omega}_{t,n}' \\
\boldsymbol{x}_{t|n} &= \boldsymbol{x}_{t|n-1} + \boldsymbol{\Omega}_{t,n}\boldsymbol{G}_n'\boldsymbol{\Delta}_n^-\left(\boldsymbol{y}_n - \boldsymbol{G}_n\hat{\boldsymbol{x}}_n\right)
\end{aligned}$$

An alternative (of course equivalent) and perhaps more appealing way to do the computation is to begin with $\boldsymbol{x}_{n|n}$ and $\boldsymbol{\Omega}_{n|n}$ from the Kalman recursions and for $t = n-1, n-2, \ldots, 1$ to compute

$$\boldsymbol{\Omega}_t^* = \boldsymbol{\Omega}_{t|t}\boldsymbol{F}_t'\boldsymbol{\Omega}_{t+1}^{-1}$$

and

$$\boldsymbol{x}_{t|n} = \boldsymbol{x}_{t|t} + \boldsymbol{\Omega}_t^*\left(\boldsymbol{x}_{t+1|n} - \hat{\boldsymbol{x}}_{t+1}\right)$$

and

$$\boldsymbol{\Omega}_{t|n} = \boldsymbol{\Omega}_{t|t} + \boldsymbol{\Omega}_t^*\left(\boldsymbol{\Omega}_{t+1|n} - \boldsymbol{\Omega}_{t+1}\right)\boldsymbol{\Omega}_t^{*\prime}$$

10.4.4 Missing Observations

Suppose that one has available observations $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{m-1}, \boldsymbol{y}_{m+1}$ (but not $\boldsymbol{y}_m$). It is still possible to make use of a version of the Kalman recursions. Note that at there is no problem in using the recursions through time $t = m-1$, producing

$$\boldsymbol{x}_{m-1|m-1} \text{ and } \boldsymbol{\Omega}_{m-1|m-1}$$

from the filtering recursion and

$$\hat{\boldsymbol{x}}_m \text{ and } \boldsymbol{\Omega}_m$$

from the prediction recursion. Then at time $t = m$ since $\boldsymbol{y}_m$ is missing, no innovation $\boldsymbol{I}_m$ can be computed. So for filtering, the usual Kalman update equation cannot be used. But (under Gaussian assumptions) one should presumably set

$$\boldsymbol{x}_{m|m} = \mathrm{E}\left[\boldsymbol{x}_m|\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{m-1}\right] = \hat{\boldsymbol{x}}_m$$

and

$$\boldsymbol{\Omega}_{m|m} = \boldsymbol{\Omega}_m$$

(both from the $t = m-1$ prediction recursion). Then for prediction, one can go ahead using

$$\hat{\boldsymbol{x}}_m = \boldsymbol{F}_m\boldsymbol{x}_{m|m}$$

and

$$\boldsymbol{\Omega}_{m+1} = \boldsymbol{F}_m\boldsymbol{\Omega}_{m|m}\boldsymbol{F}'_m + \boldsymbol{Q}_m$$

With these values, one is back on schedule to continue using the Kalman recursions beginning at $t = m+1$.

10.5 Approximately Linear State Space Modeling

A generalization of the basic state space model form replaces the linear form $\boldsymbol{F}_t \boldsymbol{x}_t$ in the state equation (54) with

$$\boldsymbol{f}_t(\boldsymbol{x}_t) = \begin{pmatrix} f_{t,1}(\boldsymbol{x}_t) \\ f_{t,2}(\boldsymbol{x}_t) \\ \vdots \\ f_{t,v}(\boldsymbol{x}_t) \end{pmatrix}$$

for some $\boldsymbol{f}_t : \Re^v \to \Re^v$ and the linear form $\boldsymbol{G}_t \boldsymbol{x}$ in the observation equation (55) with

$$\boldsymbol{g}_t(\boldsymbol{x}_t) = \begin{pmatrix} g_{t,1}(\boldsymbol{x}_t) \\ g_{t,2}(\boldsymbol{x}_t) \\ \vdots \\ g_{t,w}(\boldsymbol{x}_t) \end{pmatrix}$$

for some $\boldsymbol{g}_t : \Re^v \to \Re^w$. In the event that the $\boldsymbol{f}_t$ and $\boldsymbol{g}_t$ are differentiable functions and the state and observation error covariance matrices are relatively small (in comparison to any non-linearity in the corresponding functions), one can essentially "linearize" the model equations and use close variants of the basic Kalman equations.

That is, for

$$\dot{\boldsymbol{f}}_t(\boldsymbol{x}_0) = \begin{pmatrix} \frac{\partial}{\partial x_1} f_{t,1}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \frac{\partial}{\partial x_2} f_{t,1}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \cdots & \frac{\partial}{\partial x_v} f_{t,1}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} \\ \frac{\partial}{\partial x_1} f_{t,2}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \frac{\partial}{\partial x_2} f_{t,2}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \cdots & \frac{\partial}{\partial x_v} f_{t,2}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_{t,v}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \frac{\partial}{\partial x_2} f_{t,v}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \cdots & \frac{\partial}{\partial x_v} f_{t,v}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} \end{pmatrix}$$

and

$$\dot{\boldsymbol{g}}_t(\boldsymbol{x}_0) = \begin{pmatrix} \frac{\partial}{\partial x_1} g_{t,1}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \frac{\partial}{\partial x_2} g_{t,1}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \cdots & \frac{\partial}{\partial x_v} g_{t,1}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} \\ \frac{\partial}{\partial x_1} g_{t,2}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \frac{\partial}{\partial x_2} g_{t,2}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \cdots & \frac{\partial}{\partial x_v} g_{t,2}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} g_{t,w}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \frac{\partial}{\partial x_2} g_{t,w}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} & \cdots & \frac{\partial}{\partial x_v} g_{t,w}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0} \end{pmatrix}$$

the nonlinear state equation

$$\boldsymbol{x}_{t+1} = \boldsymbol{f}_t(\boldsymbol{x}_t) + \boldsymbol{v}_t$$

and nonlinear observation equation

$$\boldsymbol{y}_t = \boldsymbol{g}_t(\boldsymbol{x}_t) + \boldsymbol{w}_t$$

can be approximated by respectively

$$\begin{aligned} \boldsymbol{x}_{t+1} &\approx \boldsymbol{f}_t\left(\boldsymbol{x}_{t|t}\right) + \dot{\boldsymbol{f}}_t\left(\boldsymbol{x}_{t|t}\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_{t|t}\right) + \boldsymbol{v}_t \\ &= \dot{\boldsymbol{f}}_t\left(\boldsymbol{x}_{t|t}\right)\boldsymbol{x}_t + \left(\boldsymbol{f}_t\left(\boldsymbol{x}_{t|t}\right) - \dot{\boldsymbol{f}}_t\left(\boldsymbol{x}_{t|t}\right)\boldsymbol{x}_{t|t}\right) + \boldsymbol{v}_t \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{y}_t &\approx \boldsymbol{g}_t\left(\boldsymbol{x}_{t|t}\right) + \dot{\boldsymbol{g}}_t\left(\boldsymbol{x}_{t|t}\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_{t|t}\right) + \boldsymbol{w}_t \\ &= \dot{\boldsymbol{g}}_t\left(\boldsymbol{x}_{t|t}\right)\boldsymbol{x}_t + \left(\boldsymbol{g}_t\left(\boldsymbol{x}_{t|t}\right) - \dot{\boldsymbol{g}}_t\left(\boldsymbol{x}_{t|t}\right)\boldsymbol{x}_{t|t}\right) + \boldsymbol{w}_t \end{aligned}$$

These approximate model equations lead to extended Kalman recursions. Below use the abbreviations

$$\hat{\boldsymbol{F}}_t = \dot{\boldsymbol{f}}_t\left(\boldsymbol{x}_{t|t}\right) \quad \text{and} \quad \hat{\boldsymbol{G}}_t = \dot{\boldsymbol{g}}_t\left(\boldsymbol{x}_{t|t}\right)$$

Then there are

- **(Approximate) Innovation Recursions**

$$\begin{aligned} \boldsymbol{I}_t &= \boldsymbol{y}_t - \boldsymbol{g}_t\left(\hat{\boldsymbol{x}}_t\right) \quad \text{and} \\ \boldsymbol{\Delta}_t &= \hat{\boldsymbol{G}}_t\boldsymbol{\Omega}_t\hat{\boldsymbol{G}}_t' + \boldsymbol{R}_t \end{aligned}$$

- **(Approximate) Update/Filter Recursions**

$$\begin{aligned} \boldsymbol{x}_{t|t} &= \hat{\boldsymbol{x}}_t + \boldsymbol{\Omega}_t\hat{\boldsymbol{G}}_t'\boldsymbol{\Delta}_t^{-}\boldsymbol{I}_t \quad \text{and} \\ \boldsymbol{\Omega}_{t|t} &= \boldsymbol{\Omega}_t - \boldsymbol{\Omega}_t\hat{\boldsymbol{G}}_t'\boldsymbol{\Delta}_t^{-}\hat{\boldsymbol{G}}_t\boldsymbol{\Omega}_t \end{aligned}$$

  (where $\boldsymbol{\Delta}_t^{-}$ is any generalized inverse of $\boldsymbol{\Delta}_t$), and

- **(Approximate) Prediction Recursions**

$$\begin{aligned} \hat{\boldsymbol{x}}_{t+1} &= \boldsymbol{f}_t\left(\boldsymbol{x}_{t|t}\right) \quad \text{and} \\ \boldsymbol{\Omega}_{t+1} &= \hat{\boldsymbol{F}}_t\boldsymbol{\Omega}_{t|t}\hat{\boldsymbol{F}}_t' + \boldsymbol{Q}_t \end{aligned}$$

10.6 Generalized State Space Modeling, Hidden Markov Models, and Modern Bayesian Computation

One may abstract the basic structure that under Gaussian assumptions leads to the Kalman recursions. The resulting general structure, while not typically producing simple closed form prediction equations, nevertheless *is* easily handled with modern Bayesian computation software. This fact opens the possibility of quite general filtering methods. In particular, methods for time series of (not continuous but rather) discrete observations become more or less obvious. We develop these points more fully below.

Consider states and observables

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$$
$$\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$$

and adopt the notation

$$\begin{aligned} \boldsymbol{x}^t &= (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t) \quad \text{and} \\ \boldsymbol{y}^t &= (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t) \end{aligned}$$

If we consider the Gaussian version of the state space model, taken together with the initialization the state and model equations provide a fully-specified Gaussian joint distribution for the states and observables. This is built up conditionally using in succession the distributions

$$\begin{aligned} \boldsymbol{x}_1 &\sim \text{MVN}_v\left(\hat{\boldsymbol{x}}_1, \boldsymbol{\Omega}_1\right) \\ \boldsymbol{y}_1|\boldsymbol{x}_1 &\sim \text{MVN}_w\left(\boldsymbol{G}_1\boldsymbol{x}_1, \boldsymbol{R}_1\right) \\ \boldsymbol{x}_2|\boldsymbol{x}_1, \boldsymbol{y}_1 &\sim \text{MVN}_v\left(\boldsymbol{F}_1\boldsymbol{x}_1, \boldsymbol{Q}_1\right) \\ \boldsymbol{y}_2|\boldsymbol{x}^2, \boldsymbol{y}_1 &\sim \text{MVN}_w\left(\boldsymbol{G}_2\boldsymbol{x}_2, \boldsymbol{R}_2\right) \\ &\vdots \\ \boldsymbol{x}_t|\boldsymbol{x}^{t-1}, \boldsymbol{y}^{t-1} &\sim \text{MVN}_v\left(\boldsymbol{F}_{t-1}\boldsymbol{x}_{t-1}, \boldsymbol{Q}_{t-1}\right) \\ \boldsymbol{y}_t|\boldsymbol{x}^t, \boldsymbol{y}^{t-1} &\sim \text{MVN}_w\left(\boldsymbol{G}_t\boldsymbol{x}_t, \boldsymbol{R}_t\right) \\ &\vdots \\ \boldsymbol{x}_n|\boldsymbol{x}^{n-1}, \boldsymbol{y}^{n-1} &\sim \text{MVN}_v\left(\boldsymbol{F}_{n-1}\boldsymbol{x}_{n-1}, \boldsymbol{Q}_{n-1}\right) \\ \boldsymbol{y}_n|\boldsymbol{x}^n, \boldsymbol{y}^{n-1} &\sim \text{MVN}_w\left(\boldsymbol{G}_n\boldsymbol{x}_n, \boldsymbol{R}_n\right) \end{aligned}$$

Thus, for h a joint density for states and observables, f the MVN_v density, and g the MVN_w density,

$$h\left(\boldsymbol{x}^n, \boldsymbol{y}^n\right) = f\left(\boldsymbol{x}_1|\hat{\boldsymbol{x}}_1, \boldsymbol{\Omega}\right) \prod_{t=2}^{n} f\left(\boldsymbol{x}_t|\boldsymbol{F}_{t-1}\boldsymbol{x}_{t-1}, \boldsymbol{Q}_{t-1}\right) \prod_{t=1}^{n} g\left(\boldsymbol{y}_t|\boldsymbol{G}_t\boldsymbol{x}_t, \boldsymbol{R}_t\right) \quad (56)$$

This is an $n\left(v+w\right)$-dimensional Gaussian density and the Kalman recursions provide simple recursive ways of finding conditional means and conditional variances for the joint distribution. We should not expect to find such simple closed form expressions once we leave the world of Gaussian models, but the basic structure (56) does turn out to be simple enough to be handled using modern (MCMC-based) Bayes analysis software.

Since the prior mean and covariance matrix and all of the matrices in the Kalman recursions are user-supplied, the elements of the right side of display (56) are really of the forms

$$\begin{aligned} f\left(\boldsymbol{x}_1|\hat{\boldsymbol{x}}_1, \boldsymbol{\Omega}\right) &= f_1\left(\boldsymbol{x}_1\right) \\ f\left(\boldsymbol{x}_t|\boldsymbol{F}_{t-1}\boldsymbol{x}_{t-1}, \boldsymbol{Q}_{t-1}\right) &= f_t\left(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}\right) \quad \text{and} \\ g\left(\boldsymbol{y}_t|\boldsymbol{G}_t\boldsymbol{x}_t, \boldsymbol{R}_t\right) &= g_t\left(\boldsymbol{y}_t|\boldsymbol{x}_t\right) \end{aligned}$$

for a particular (user-supplied) density f_1 and user supplied conditional densities f_t and g_t. Using these notations, a joint probability structure motivated by the Gaussian version of the state space model has density

$$h\left(\boldsymbol{x}^n, \boldsymbol{y}^n\right) = \left[f_1\left(\boldsymbol{x}_1\right) \prod_{t=2}^{n} f_t\left(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}\right)\right] \prod_{t=1}^{n} g_t\left(\boldsymbol{y}_t | \boldsymbol{x}_t\right) \tag{57}$$

the bracketed part of which specifies a Markov chain model for the states. Conditioned on the states, the observations are independent, $\boldsymbol{y}_t$ with a distribution depending upon $\boldsymbol{x}^n$ only through $\boldsymbol{x}_t$. As one only sees states through observations, form (57) can appropriately be called a "hidden Markov model" or "generalized state space model."

Now, again, form (57) will in general not provide simple closed forms for the conditional distributions of observations of $\boldsymbol{x}$'s and $\boldsymbol{y}$'s that provide filters and predictions. But that is more or less irrelevant in the modern computing environment. Form (57) is easily programmed into modern Bayes software, and for $\boldsymbol{y}^*$ any subset of $\boldsymbol{y}^n$, MCMC-based simulations then provide (posterior) conditional distributions for all of the entries of $\boldsymbol{x}^n$ and of $\boldsymbol{y}^n - \boldsymbol{y}^*$ (any unobserved/missing "observables"). This is incredibly powerful.

In fact, even more is possible. The form (57) can be generalized to

$$h\left(\boldsymbol{x}^n, \boldsymbol{y}^n | \boldsymbol{\Lambda}\right) = \left[f_1\left(\boldsymbol{x}_1 | \boldsymbol{\Lambda}\right) \prod_{t=2}^{n} f_t\left(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, \boldsymbol{\Lambda}\right)\right] \prod_{t=1}^{n} g_t\left(\boldsymbol{y}_t | \boldsymbol{x}_t, \boldsymbol{\Lambda}\right)$$

for a vector parameter $\boldsymbol{\Lambda}$, and upon providing a distribution for $\boldsymbol{\Lambda}$ through a density $k\left(\boldsymbol{\Lambda}\right)$, the "hierarchical" form

$$k\left(\boldsymbol{\Lambda}\right) h\left(\boldsymbol{x}^n, \boldsymbol{y}^n | \boldsymbol{\Lambda}\right)$$

is equally easily entered and used in software like `OpenBUGS/WinBUGS`, producing filtered values, predictions and corresponding uncertainties.

10.7 State Space Representations of ARIMA Models

As indicated in Section 10.1, general ARIMA models have state space representations. We proceed to provide those. The place to begin is with causal AR(p) models. The standard representation of scalar values from such a process is of course

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

We consider the p-dimensional state variable

$$\boldsymbol{x}_t = \begin{pmatrix} y_{t-p+1} \\ y_{t-p+2} \\ \vdots \\ y_t \end{pmatrix}$$

and observation equation

$$y_t = \underset{p\times 1}{(0,0,\ldots,0,1)}\boldsymbol{x}_t + 0$$

An appropriate state equation is then

$$\boldsymbol{x}_{t+1} = \begin{bmatrix} 0 & & & \\ \vdots & & \underset{(p-1)\times(p-1)}{\boldsymbol{I}} & \\ 0 & & & \\ \phi_p & \phi_{p-1} & \cdots & \phi_1 \end{bmatrix} \boldsymbol{x}_t + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \epsilon_{t+1}$$

and then (with a proper initialization) the AR(p) model has a state space representation with

$$\boldsymbol{G} = (0,0,\ldots,0,1)\,, w_t = 0, \boldsymbol{F} = \begin{bmatrix} 0 & & & \\ \vdots & & \underset{(p-1)\times(p-1)}{\boldsymbol{I}} & \\ 0 & & & \\ \phi_p & \phi_{p-1} & \cdots & \phi_1 \end{bmatrix}, \text{ and } \boldsymbol{v}_t = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \epsilon_t \end{pmatrix}$$

(Personally, I would worry little about getting the initialization that makes the state space representation of $\boldsymbol{Y}$ exactly second order stationary, expecting that typically any sane initialization would produce about the same forecasts beyond time n for practical values of n.)

Next, consider the problem of representing a causal ARMA(p,q) process. Consider the basic ARMA equation in operator form

$$\Phi(\mathcal{B})\,\boldsymbol{Y} = \Theta(\mathcal{B})\,\boldsymbol{\epsilon}$$

and let $r = \max(p, q+1)$ so that $\phi_j = 0$ for $j \geq r$, $\theta_0 = 1$, and $\theta_j = 0$ for $j \geq r$. If $\boldsymbol{U}$ is the causal AR(p) process satisfying

$$\Phi(\mathcal{B})\,\boldsymbol{U} = \boldsymbol{\epsilon}$$

then

$$\Phi(\mathcal{B})\,\boldsymbol{Y} = \Theta(\mathcal{B})\,\Phi(\mathcal{B})\,\boldsymbol{U} = \Phi(\mathcal{B})\,\Theta(\mathcal{B})\,\boldsymbol{U}$$

$\Phi(\mathcal{B})$ is invertible and

$$(\Phi(\mathcal{B}))^{-1}\,\Phi(\mathcal{B})\,\boldsymbol{Y} = (\Phi(\mathcal{B}))^{-1}\,\Phi(\mathcal{B})\,\Theta(\mathcal{B})\,\boldsymbol{U}$$

and thus

$$\boldsymbol{Y} = \Theta(\mathcal{B})\,\boldsymbol{U}$$

So for

$$\boldsymbol{x}_t = \begin{pmatrix} u_{t-r+1} \\ u_{t-r+2} \\ \vdots \\ u_t \end{pmatrix}$$

one can write

$$y_t = (\theta_{r-1}, \theta_{r-2}, \ldots, \theta_1, \theta_0)\, \boldsymbol{x}_t + 0$$

(an observation equation) where from the AR(p) case, we can write a state equation as

$$\boldsymbol{x}_{t+1} = \begin{bmatrix} 0 & & & \\ \vdots & & \underset{(r-1)\times(r-1)}{\boldsymbol{I}} & \\ 0 & & & \\ \phi_r & \phi_{r-1} & \cdots & \phi_1 \end{bmatrix} \boldsymbol{x}_t + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \epsilon_{t+1} \tag{58}$$

So (with a proper initialization) the ARMA(p, q) model has state space representation with

$$\boldsymbol{G} = (\theta_{r-1}, \theta_{r-2}, \ldots, \theta_1, \theta_0)\, , w_t = 0, \tag{59}$$

$$\boldsymbol{F} = \begin{bmatrix} 0 & & & \\ \vdots & & \underset{(r-1)\times(r-1)}{\boldsymbol{I}} & \\ 0 & & & \\ \phi_r & \phi_{r-1} & \cdots & \phi_1 \end{bmatrix}, \text{ and } \boldsymbol{v}_t = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \epsilon_t \end{pmatrix} \tag{60}$$

(and again, I personally would not much concern myself with identifying the initialization that makes the $\boldsymbol{Y}$ model exactly second order stationary).

So then consider developing a state space representation of an ARIMA(p, d, q) process model. Suppose that $\boldsymbol{Z} = \mathcal{D}^d \boldsymbol{Y}$ is ARMA(p, q), satisfying

$$\Phi(\mathcal{B})\, \boldsymbol{Z} = \Theta(\mathcal{B})\, \boldsymbol{\epsilon}$$

Then applying the ARMA development above to $\boldsymbol{Z}$, we have a state space representation with observation equation

$$z_t = (\theta_{r-1}, \theta_{r-2}, \ldots, \theta_1, \theta_0)\, \boldsymbol{x}_t + 0$$

and state equation (58).

Note then that

$$\begin{aligned} \mathcal{D}^d \boldsymbol{Y} &= (\mathcal{I} - \mathcal{B})^d\, \boldsymbol{Y} \\ &= \left(\sum_{j=0}^{d} (-1)^j\, \mathcal{B}^j \mathcal{I}^{d-j} \right) \boldsymbol{Y} \\ &= \sum_{j=0}^{d} (-1)^j\, \mathcal{B}^j\, \boldsymbol{Y} \end{aligned}$$

so that

$$\boldsymbol{Y} = \mathcal{D}^d \boldsymbol{Y} - \sum_{j=1}^{d} (-1)^j\, \mathcal{B}^j\, \boldsymbol{Y} = \boldsymbol{Z} - \sum_{j=1}^{d} (-1)^j\, \mathcal{B}^j\, \boldsymbol{Y}$$

Thus, with

$$\boldsymbol{y}_t = \begin{pmatrix} y_{t-d+1} \\ y_{t-d+2} \\ \vdots \\ y_t \end{pmatrix}$$

and

$$\underset{d\times 1}{\boldsymbol{A}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \text{ and } \underset{d\times d}{\boldsymbol{B}} = \begin{bmatrix} 0 & & & \\ \vdots & & \underset{(d-1)\times(d-1)}{\boldsymbol{I}} & \\ 0 & & & \\ (-1)^{d+1}\binom{d}{d} & (-1)^{d}\binom{d}{d-1} & \cdots & d \end{bmatrix} \tag{61}$$

it is the case that

$$y_t = \boldsymbol{A}\boldsymbol{z}_t + \boldsymbol{B}\boldsymbol{y}_{t-1} = \boldsymbol{A}\boldsymbol{G}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{y}_{t-1}$$

for $\boldsymbol{G}$ as in display (59) and $\boldsymbol{A}$ and $\boldsymbol{B}$ as in display (61). So with $\boldsymbol{v}_t$ as in display (60), defining a new state vector and state error

$$\underset{(r+d)\times 1}{\boldsymbol{x}_t^*} = \begin{pmatrix} \boldsymbol{x}_t \\ \boldsymbol{y}_{t-1} \end{pmatrix} \text{ and } \underset{(r+d)\times 1}{\boldsymbol{v}_t^*} = \begin{pmatrix} \boldsymbol{v}_t \\ \boldsymbol{0} \end{pmatrix}$$

the new state equation

$$\boldsymbol{x}_{t+1}^* = \begin{pmatrix} \boldsymbol{F} & \boldsymbol{0} \\ \boldsymbol{A}\boldsymbol{G} & \boldsymbol{B} \end{pmatrix} \boldsymbol{x}_t^* + \boldsymbol{v}_t^*$$

(with $\boldsymbol{F}$ as in display (60)) and observation equation

$$y_t = \left(\boldsymbol{G}, (-1)^{d+1}\binom{d}{d}, (-1)^{d}\binom{d}{d-1}, \ldots, d\right)\boldsymbol{x}_t^* + 0$$

provide a state space representation of an ARIMA(p,d,q) model with

$$\boldsymbol{G}^* = \left(\boldsymbol{G}, (-1)^{d+1}\binom{d}{d}, (-1)^{d}\binom{d}{d-1}, \ldots, d\right), w_t = 0, \text{ and } \boldsymbol{F}^* = \begin{pmatrix} \boldsymbol{F} & \boldsymbol{0} \\ \boldsymbol{A}\boldsymbol{G} & \boldsymbol{B} \end{pmatrix}$$

(This, of course, is technically subject to using an initialization that makes the $\boldsymbol{Y}$ process exactly second order stationary. But again, I personally don't see this as a serious practical issue.)

Note that in the event that one wishes to represent a "subset ARIMA" model in state space form, all that is required is to set appropriate ϕ_j's in $\boldsymbol{F}$ and/or θ_j's in $\boldsymbol{G}$ equal to 0.

The whole development just concluded for ARIMA(p,d,q) models has exact parallels for other cases of differencing of $\boldsymbol{Y}$. Consider, for a concrete example, the case where $\mathcal{D}^* = \mathcal{D}\mathcal{D}_4$ and one wishes to model $\mathcal{D}^*\boldsymbol{Y}$ as ARMA(p,q). Since

$$\mathcal{D}\mathcal{D}_4 = \mathcal{I} - \mathcal{B} - \mathcal{B}^4 + \mathcal{B}^5$$

it follows that

$$\boldsymbol{Y} = \mathcal{D}^* \boldsymbol{Y} + \mathcal{B}\boldsymbol{Y} + \mathcal{B}^4 \boldsymbol{Y} - \mathcal{B}^5 \boldsymbol{Y}$$

If we then set

$$\boldsymbol{y}_t = \begin{pmatrix} y_{t-4} \\ y_{t-3} \\ \vdots \\ y_t \end{pmatrix}, \underset{5\times 1}{\boldsymbol{A}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \text{ and } \underset{5\times 5}{\boldsymbol{B}} = \begin{bmatrix} 0 & & & & \\ 0 & & & & \\ 0 & & \underset{4\times 4}{\boldsymbol{I}} & & \\ 0 & & & & \\ -1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

we have as for the ARIMA(p, d, q) case

$$y_t = \boldsymbol{A}\boldsymbol{z}_t + \boldsymbol{B}\boldsymbol{y}_{t-1}$$

and may proceed as before.

11 "Other" Time Series Models

We consider some less standard/less widely used time series models.

11.1 ARCH and GARCH Models for Describing Conditional Heteroscedasticity

For ARIMA models, the conditional variance of $y_t | y_{t-1}, y_{t-2}, y_{t-3} \ldots$ is constant (does not depend upon t or the values of past observations). Financial time series (for example for the log ratios of closing prices of a stock on successive trading days) often exhibit what seem to be non-constant conditional variances. These seem to be big where the immediate past few values of the y_t series are varying wildly and to be small where the immediate past few values are relatively similar. That is, such series exhibit "volatility clustering." So called "ARCH" (autoregressive conditionally heteroscedastic) models and "GARCH" (generalized ARCH) models have been proposed to represent this behavior.

11.1.1 Modeling

For $\alpha_0 > 0$ and $0 < \alpha_j < 1$ for $j = 1, 2, \ldots, p$ use the notation

$$h_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j y_{t-j}^2 \tag{62}$$

and then for an iid N$(0, 1)$ sequence of variables $\{\epsilon_t\}$, the variables

$$y_t = \sqrt{h_t}\epsilon_t \tag{63}$$

are said to have an ARCH(p) joint distribution. It's easy to argue that under this model

$$\text{Var}\left(y_t | y_{t-1}, y_{t-2}, y_{t-3} \ldots\right) = h_t$$

which is obviously non-constant in the previous observations, having a floor value of α_0 and increasing in the volatility of the (p) immediately preceding observations. Further, ϵ_t is independent of $y_{t-1}, y_{t-2}, y_{t-3}\ldots$, from which it's easy to see that $\mathrm{E}y_t = 0$. In fact, it turns out that $\boldsymbol{Y}$ is strictly (and therefore second order) stationary. The variance of the process, $\sigma^2 = \mathrm{Var}y_t$, may be derived as

$$\begin{aligned}\mathrm{Var}y_t &= \mathrm{EVar}\left(y_t|y_{t-1}, y_{t-2}, y_{t-3}\ldots\right) + \mathrm{Var}\left(\mathrm{E}\left[y_t|y_{t-1}, y_{t-2}, y_{t-3}\ldots\right]\right)\\ &= \mathrm{EVar}\left(y_t|y_{t-1}, y_{t-2}, y_{t-3}\ldots\right) + 0\end{aligned}$$

i.e.

$$\mathrm{E}y_t^2 = \mathrm{E}\left(\alpha_0 + \sum_{j=1}^{p}\alpha_j y_{t-j}^2\right)$$

so that

$$\sigma^2 = \alpha_0 + \sigma^2\sum_{j=1}^{p}\alpha_j$$

and thus

$$\sigma^2 = \frac{\alpha_0}{1 - \sum_{j=1}^{p}\alpha_j}$$

(Notice, by the way, that this result indicates the necessity of $\sum_{j=1}^{p}\alpha_j < 1$ in an ARCH model.)

Notice also that if we define

$$\eta_t = y_t^2 - h_t$$

it is the case that

$$\eta_t = \epsilon_t^2 h_t - h_t = h_t\left(\epsilon_t^2 - 1\right)$$

The series $\{\eta_t\}$ then clearly has mean 0 and constant variance. As it turns out, it is also uncorrelated and is thus a white noise series. So (from the definition of η_t)

$$\begin{aligned}y_t^2 &= h_t + \eta_t\\ &= \alpha_0 + \sum_{j=1}^{p}\alpha_j y_{t-j}^2 + \eta_t\end{aligned} \tag{64}$$

and we see that $\left\{y_t^2\right\}$ is an AR(p) series.

One way that ARCH models have been generalized is to make GARCH(p, q) models where one assumes that the basic relationship (63) holds, but the form (62) is generalized to

$$h_t = \alpha_0 + \sum_{j=1}^{p}\alpha_j y_{t-j}^2 + \sum_{j=1}^{q}\beta_j h_{t-j}$$

where $\alpha_0 > 0$ and each $\alpha_j > 0$ for $j \geq 1$ and each $\beta_j > 0$ for $j \geq 1$. Here conditional variances depend upon immediately preceding observations and immediately preceding conditional variances.

11.1.2 Inference for ARCH Models

In the basic ARCH model, $\sqrt{h_t}$ is a scaling factor that depends only upon past y's in the generation of y_t. If we let $f(z)$ be any fixed pdf (including especially the standard normal pdf, $f(z) = \phi(z)$) and for $h > 0$ take

$$f(y|h) = \frac{1}{\sqrt{h}} f\left(\frac{y}{\sqrt{h}}\right)$$

to be the scaled version of f (so that Z having density f means that $\sqrt{h}Z$ has density $f(\cdot|h)$). A version of the ARCH model says that the joint density for $y_{p+1}, y_{p+2}, \ldots, y_n$ conditioned on $\boldsymbol{Y}_p$ and depending upon the set of ARCH parameters, $\boldsymbol{\alpha}$, is

$$f(y_{p+1}, y_{p+2}, \ldots, y_n | y_1, y_2, \ldots, y_p, \boldsymbol{\alpha}) = \prod_{t=p+1}^{n} f(y_t|h_t)$$

and for that matter, the joint density for $\boldsymbol{Y}_n$ conditioned on (unobservable) $y_{-p+1}, y_{-p+2}, \ldots, y_0$ and depending upon the set of ARCH parameters, $\boldsymbol{\alpha}$, is

$$f(y_1, y_2, \ldots, y_n | y_{-p+1}, y_{-p+2}, \ldots, y_0, \boldsymbol{\alpha}) = \prod_{t=1}^{n} f(y_t|h_t)$$

In both of these expressions dependence upon $\boldsymbol{\alpha}$ enters the right hand side through the factors, h_t, that are variances in the normal case. The generalization here beyond the normal case opens the possibility of using "heavy-tailed" densities $f(z)$ (like t densities), a development that seems to be of some importance in financial applications of these models.

In any event, one method of inference/estimation in ARCH models is to use

$$L(\boldsymbol{\alpha}) = \ln f(y_{p+1}, y_{p+2}, \ldots, y_n | y_1, y_2, \ldots, y_p, \boldsymbol{\alpha})$$

as a conditional log-likelihood. Maximization of $L(\boldsymbol{\alpha})$ then produces conditional MLE's for the ARCH parameters, and use of the Hessian matrix evaluated at the MLE leads to an estimated covariance matrix for MLE's of the individual α_j's and thus standard errors of estimation. Notice that at least as developed thus far, plugging estimated parameters into the model, point predictions of future observations are all 0, and simulation from existing observables into the future can provide prediction variances.

An alternative to use of a conditional likelihood might be to employ an approximate unconditional likelihood produced as follows. In light of the recursion

(64) let

$$\begin{aligned}
\widehat{y_0^2} &= \frac{1}{\alpha_p}\left(y_p^2 - \alpha_0 - \alpha_1 y_{p-1}^2 - \alpha_2 y_{p-2}^2 - \cdots - \alpha_{p-1} y_1^2\right) \\
\widehat{y_{-1}^2} &= \frac{1}{\alpha_p}\left(y_{p-1}^2 - \alpha_0 - \alpha_1 y_{p-2}^2 - \alpha_2 y_{p-3}^2 - \cdots - \alpha_{p-2} y_1^2 - \alpha_{p-1}\widehat{y_0^2}\right) \\
\widehat{y_{-2}^2} &= \frac{1}{\alpha_p}\left(y_{p-2}^2 - \alpha_0 - \alpha_1 y_{p-3}^2 - \alpha_2 y_{p-4}^2 - \cdots - \alpha_{p-3} y_1^2 - \alpha_{p-2}\widehat{y_0^2} - \alpha_{p-1}\widehat{y_{-1}^2}\right) \\
&\vdots \\
\widehat{y_{-p+1}^2} &= \frac{1}{\alpha_p}\left(y_1^2 - \alpha_0 - \alpha_1 \widehat{y_0^2} - \alpha_2 \widehat{y_{-1}^2} - \cdots - \alpha_{p-2}\widehat{y_{-p+3}^2} - \alpha_{p-1}\widehat{y_{-p+2}^2}\right)
\end{aligned}$$

(this amounts to "back-casting" p squared observations based on the AR(p) model for these squares). Then define for $1 \leq t \leq p$

$$\widehat{h_t} = \alpha_0 + \sum_{j=1}^{t-1} \alpha_j y_{t-j}^2 + \sum_{j=t-1}^{p} \alpha_j \widehat{y_{t-j}^2}$$

and in place of the conditional log likelihood one might instead use

$$L^*(\boldsymbol{\alpha}) = \sum_{t=1}^{p} \ln f\left(y_t | \widehat{h_t}\right) + \sum_{t=p+1}^{n} \ln f\left(y_t | h_t\right)$$

What strikes me as another more interesting and potentially more effective method of inference (providing coherent predictions and even handling of missing values most directly) is to use modern Bayes computing. That is, using the conditional model for $\boldsymbol{Y}_{n+s}$ provided by $f\left(y_1, y_2, \ldots, y_{n+s} | y_{-p+1}, y_{-p+2}, \ldots, y_0, \boldsymbol{\alpha}\right)$ and some kind of prior distributions for $y_{-p+1}, y_{-p+2}, \ldots, y_0, \boldsymbol{\alpha}$, say specified by $g\left(y_{-p+1}, y_{-p+2}, \ldots, y_0, \boldsymbol{\alpha}\right)$, one has a joint distribution for all variables specified by

$$f\left(y_1, y_2, \ldots, y_{n+s} | y_{-p+1}, y_{-p+2}, \ldots, y_0, \boldsymbol{\alpha}\right) g\left(y_{-p+1}, y_{-p+2}, \ldots, y_0, \boldsymbol{\alpha}\right)$$

where the first term is very easily coded in software like `WinBUGS/OpenBUGS`. Then upon plugging in observed values of some subset of $y_1, y_2, \ldots, y_n$, simulated conditional distributions of the remaining entries of $\boldsymbol{Y}_{n+s}$, future values $y_{n+1}, y_{n+2}, \ldots, y_{n+s}$, and parameters in $\boldsymbol{\alpha}$ provide filtering, prediction, and estimation in this Bayes model.

What to use for a prior distribution for $y_{-p+1}, y_{-p+2}, \ldots, y_0, \boldsymbol{\alpha}$ is not completely obvious, but here is what I might try first. Recognizing that one must have each $\alpha_j > 0$ and $\sum_{j=1}^{p} \alpha_j < 1$ in an ARCH model and (at least in the normal model) that α_0 is a minimal conditional variance, I might try making *a priori*

$$\ln \alpha_0 \sim \mathrm{U}\left(-\infty, \infty\right)$$

or perhaps

$$\sqrt{\alpha_0} \sim \mathrm{U}\left(0, \infty\right)$$

independent of

$$(\alpha_1, \alpha_2, \ldots \alpha_p, \gamma) \sim \text{Dirichlet}_{p+1}\left(\left(\frac{1}{p+1}, \frac{1}{p+1}, \ldots \frac{1}{p+1}\right)\right)$$

(for a variable γ that never really gets used in the model). What should then get used as a conditional distribution for $y_{-p+1}, y_{-p+2}, \ldots, y_0$ given $\boldsymbol{\alpha}$ is not obvious. One possibility that I might try is this. First, in view of form of the variance for the ARCH model, one might assume that *a priori*

$$y_{-p+1}|\boldsymbol{\alpha} \sim f\left(y \left| \frac{\alpha_0}{1 - \sum_{j=1}^{p} \alpha_j} \right.\right)$$

and then in succession for $-p+1 < t \leq 0$, that *a priori*

$$y_t|\boldsymbol{\alpha}, y_{-p+1}, y_{-p+2}, \ldots, y_{t-1} \sim f\left(y \left| \alpha_0 + \sum_{j=1}^{t+p-1} \alpha_j y_{t-j}^2 + \left(\frac{\alpha_0}{1 - \sum_{j=1}^{p} \alpha_j}\right) \sum_{j=t+p}^{p} \alpha_j \right.\right)$$

Another possibility is to simply ignore the dependence in $(y_{-p+1}, y_{-p+2}, \ldots, y_0)$ and set

$$(y_{-p+1}, y_{-p+2}, \ldots, y_0)\,|\boldsymbol{\alpha}, \gamma \sim \text{MVN}_p\left(\mathbf{0}, B\boldsymbol{I}\right)$$

for a constant $B \geq \alpha_0 / \left(1 - \sum_{j=1}^{p} \alpha_j\right)$.

11.2 Self-Exciting Threshold Auto-Regressive Models

The basic idea here is that the auto-regressive model that a process follows in the generation of y_t depends upon the immediate past value of the process, y_{t-1}. For simplicity, we'll describe a simple two-regime version of the modeling here, but extension to more than two regimes is more or less obvious/straightforward.

Consider two sets of parameters for AR(p) models

$$\begin{aligned} &\phi_{01}, \phi_{11}, \ldots, \phi_{p1}, \sigma_1 \quad \text{and} \\ &\phi_{02}, \phi_{12}, \ldots, \phi_{p2}, \sigma_2 \end{aligned}$$

Then for $\{\epsilon_t\}$ iid random variables with mean 0 and standard deviation 1, suppose that

$$y_t = \begin{cases} \phi_{01} + \phi_{11} y_{t-1} + \cdots + \phi_{p1} y_{t-p} + \sigma_1 \epsilon_t & \text{if } \; y_t \leq r \\ \phi_{02} + \phi_{12} y_{t-1} + \cdots + \phi_{p2} y_{t-p} + \sigma_2 \epsilon_t & \text{if } \; y_t > r \end{cases}$$

Here both the conditional mean and the conditional standard deviation of $y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ depend upon how the value y_{t-1} compares to some threshold, r.

For f the marginal density of the errors ϵ, let

$$f\left(y|\mu, \sigma\right) = \frac{1}{\sigma} f\left(\frac{y - \mu}{\sigma}\right)$$

Then (suppressing dependence upon $\boldsymbol{\phi}_1, \sigma_1, \boldsymbol{\phi}_2, \sigma_2, r$ on the left side of the equation) the conditional density of $y_t|\boldsymbol{Y}_{t-1}$ is

$$\begin{aligned} g\left(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}\right) &= I\left[y_{t-1} \leq r\right] f\left(y_t|\phi_{01} + \phi_{11}y_{t-1} + \cdots + \phi_{p1}y_{t-p}, \sigma_1\right) \\ &\quad + I\left[y_{t-1} > r\right] f\left(y_t|\phi_{02} + \phi_{12}y_{t-1} + \cdots + \phi_{p2}y_{t-p}, \sigma_2\right) \end{aligned}$$

So the conditional density of $y_{p+1}, \ldots, y_n|\boldsymbol{Y}_p$ is

$$\prod_{t=p+1}^{n} g\left(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}, \boldsymbol{\phi}_1, \sigma_1, \boldsymbol{\phi}_2, \sigma_2, r\right)$$

where we are now displaying dependence upon $\boldsymbol{\phi}_1, \sigma_1, \boldsymbol{\phi}_2, \sigma_2, r$. This leads to a conditional log-likelihood

$$L\left(\boldsymbol{\phi}_1, \sigma_1, \boldsymbol{\phi}_2, \sigma_2, r\right) = \sum_{t=p+1}^{n} \ln g\left(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}, \boldsymbol{\phi}_1, \sigma_1, \boldsymbol{\phi}_2, \sigma_2, r\right)$$

that can be used as a basis of inference more or less as for ARCH/GARCH models. (One point that is worth noticing here is that L is piecewise constant in r, jumping only at observed values of the elements of $y_{p+1}, \ldots, y_p$. The model is thus not "regular" and some alternative to the simple use of a Hessian matrix must be employed to find standard errors for the parameter estimates.) Or, one could treat (unobserved) values $y_0, y_1, \ldots, y_{-p+1}$ as part of the modeling, set prior distributions on all of $y_0, y_1, \ldots, y_{-p+1}, \boldsymbol{\phi}_1, \sigma_1, \boldsymbol{\phi}_2, \sigma_2, r$, and use modern Bayes MCMC software to enable inference. Once one identifies sensible priors, this approach has the advantages of more or less automatically handling the non-regularity of the data model, missing values in the series, and prediction beyond time n.