Vector Autoregressions
and Their Use in the Validation
of Proposed Leading Indicators

Tom Fomby
Department of Economics
Southern Methodist University
Dallas, TX 75275
September 2010

I. Introduction

In the past we have focused on the building of accurate univariate time series forecasting models (Box-Jenkins, UCM, etc.). These models rely solely on various characterizations of the trend, seasonality, and cycle of the time series being modeled (the target variable). But what if there exist other variables, supplementary variables, which, when modeled in conjunction with the target variable, provide more accurate forecasts of the target variable. A very popular multiple time series model that incorporates these supplementary variables is called the Vector Autoregressive (VAR) model. Here we will use this model to judge the usefulness of proposed supplementary (leading indicator) economic and business variables, say x, w, and z, for improving the forecasting accuracy of target variables which we label y.

For example, consider the two time series y and x plotted below by the SAS program Mplot.sas. These two series are taken from the Box and Jenkins (1970) textbook. They call the series the "M" data set. The first graph plots the raw data while the second graph plots the 3-term, centered moving average of the series. Taking the moving averages of the series makes them smoother and by so doing it is then easier to see the relationship between the series. Notice that the peaks and troughs of the x series lead the peaks and troughs of the y series by roughly 3 months. The question then is whether we can forecast y more accurately by using the "supplemental" information that seems to be present in the x series. That is the task of this chapter.

The source of the M data series is not given in the Box and Jenkins textbook and therefore may be simulated series. Regardless, they model the two time series simultaneously using a so-called transfer function model. We use a slightly different approach, the VAR model, in modeling the two time series together. To make this presentation more motivating, let us assume the y series represents the monthly sales in the SHAZAM carpet store in Flower Mound, Texas. Furthermore, assume the x series is a monthly series of observations on the housing starts in the county where Flower Mound is located. Obviously, housing starts eventually lead to housing completions and homes sold with extensive carpeting in them. Then, at least in theory, the housing starts series could be helpful in focusing SHAZAM's carpet sales.

To examine the efficacy of the x series as a useful supplement in forecasting the y series, we can conduct an **out-of-sample forecasting experiment**. Let us partition the 150 observations on these series into two parts – the first 120 observations being designated as the **in-sample data set** and the last 30 observations being designated as **the out-of-sample data set**. To carry out the experiment we need to carry out a "forecasting competition" between a purely statistical model for the y series like the Box-Jenkins model and a model that, in addition to modeling the target series (y), also incorporates the potentially useful supplemental (leading) indicator (x). This model is the Vector Autoregressive (VAR) model. Then, given the out-of-sample experiment, if the VAR model, which incorporates the leading indicator, produces more accurate forecasts of the target variable than a univariate Box-Jenkins model that ignores the potential usefulness of the x series, we can conclude that the leading indicator is useful in forecasting y.

II. The Vector Autoregressive Model

Before we turn to examining the efficacy of the x series as a leading indicator of y in the M data set, let us establish the notation and particulars of the Vector Autoregressive model. Let $y_{t1} = y_t$ and $y_{t2} = x_t$ be time series observations at time t on the target variable and leading indicator series, respectively. Then the first-order Vector Autoregressive model, abbreviated VAR(1), is written as

$$\begin{aligned} y_{t1} &= \emptyset_{10} + \emptyset_{11} y_{t-1,1} + \emptyset_{12} y_{t-1,2} + a_{t1} \\ y_{t2} &= \emptyset_{20} + \emptyset_{21} y_{t-1,1} + \emptyset_{22} y_{t-1,2} + a_{t2} \ . \end{aligned} \tag{1}$$

The first equation of the system (1) specifies that the first variable is a function of one lagged value of itself and one lagged value of the second variable while the second equation specifies that the second variable is a function of one lagged value of itself and one lagged value of the first variable. The errors of the equations (a_{t1} and a_{t2}) are assumed to have zero means and to be uncorrelated among themselves and only contemporaneously correlated between each other. In particular the errors satisfy the following properties.

$$E\begin{pmatrix} a_{t1} \\ a_{t2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{for all t} \tag{2}$$

$$E\begin{pmatrix} a_{s1}a_{t1} & a_{s1}a_{t2} \\ a_{s2}a_{t1} & a_{s2}a_{t2} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \text{ for all } s \neq t \tag{3}$$

$$\mathrm{E}\begin{pmatrix} a_{s1}a_{t1} & a_{s1}a_{t2} \\ a_{s2}a_{t1} & a_{s2}a_{t2} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \text{ for all } s = t \tag{4}$$

and $\sigma_{12} = \sigma_{21}$ in equation (4). Loosely speaking, we refer to these errors as being "white noise" errors.

As shown by Sims (1980), the equation system (1) can be thought of as a unrestricted reduced form of a simultaneous equation system involving two endogenous variables $y_{t1}$and y_{t2}. A long history of the usage of this model and its p-lag length generalization has shown it to be useful as a forecasting method. See, for example, (references). As the length of the lags on each of the variables in each of the equations is the same (one), the VAR system is called an **equal-lag length VAR**.

The p-equal-length VAR generalization of (1) is given by the system (5) below.

$$\begin{pmatrix} y_{t1} \\ y_{t2} \end{pmatrix} = \begin{pmatrix} \emptyset_{10} \\ \emptyset_{20} \end{pmatrix} + \begin{pmatrix} \emptyset_{11}^{(1)} & \emptyset_{12}^{(1)} \\ \emptyset_{21}^{(1)} & \emptyset_{22}^{(1)} \end{pmatrix} \begin{pmatrix} y_{t-1,1} \\ y_{t-1,2} \end{pmatrix} + \begin{pmatrix} \emptyset_{11}^{(2)} & \emptyset_{12}^{(2)} \\ \emptyset_{21}^{(2)} & \emptyset_{22}^{(2)} \end{pmatrix} \begin{pmatrix} y_{t-2,1} \\ y_{t-2,2} \end{pmatrix}$$

$$+ \cdots + \begin{pmatrix} \emptyset_{11}^{(p)} & \emptyset_{12}^{(p)} \\ \emptyset_{21}^{(p)} & \emptyset_{22}^{(p)} \end{pmatrix} \begin{pmatrix} y_{t-p,1} \\ y_{t-p,2} \end{pmatrix} + \begin{pmatrix} a_{t1} \\ a_{t2} \end{pmatrix}. \qquad (5)$$

The matrix form of (5) is given by

$$\mathbf{y_t} = \boldsymbol{\emptyset_0} + \boldsymbol{\emptyset_1}\mathbf{y_{t-1}} + \boldsymbol{\emptyset_2}\mathbf{y_{t-2}} + \cdots + \boldsymbol{\emptyset_p}\mathbf{y_{t-p}} + \mathbf{a}_t \ . \qquad (6)$$

In estimating the coefficients of (5) it is assumed that the two series y_{t1} and y_{t2} have been transformed to stationary form (constant mean, constant variance, and constant covariance) by differencing the original series or possibly transforming the data to percentage-change form by taking the natural log of the data and then differencing the logged data as in $z_t = ln(x_t) - \ln(x_{t-1})$ when the data has an exponential growth appearance. Of course, if the original data is already stationary, then no transformation of the series is needed. In the case of the series M data, Augmented Dickey-Fuller tests indicate that both series need to be differenced in order to make them stationary. Thus, here $y_{t1} = \Delta y_t$ and $y_{t2} = \Delta x_t$. A sufficient condition for the VAR(p) system to be stationary is that all of the eigenvalues of the following “companion” matrix be less than one in modulus (i.e. the absolute value of a real root is less than one and when the root is complex as in a $\pm$ bi, then $\sqrt{a^2 + b^2} < 1.$)

$$\begin{pmatrix} \emptyset_1 & \emptyset_2 & \cdots & \emptyset_p \\ I_2 & 0 & \cdots & 0 \\ 0 & I_2 & 0 & \cdots \\ \vdots & \ddots & \vdots & 0 \\ 0 & \cdots & I_2 & 0 \end{pmatrix} \qquad (7)$$

The matrix (7) is of dimension 2p x 2p. If, when the coefficients in (7) are replaced by their estimates, the eigenvalues of such an estimated matrix are not all less than one in modulus, then one might think again as to whether the right dimension (p) has been determined for the VAR or possibly that, contrary to first opinion, the variables of the VAR have not been sufficiently transformed to stationarity.

III. Choosing the Right Lag-Length for the VAR