

# The Econometric Analysis of Some Constructed Binary Time Series\*

Don Harding<sup>†</sup> and Adrian Pagan<sup>‡</sup>

23 March 2007

## Abstract

Macroeconometric and financial researchers often use *secondary* or *constructed* binary random variables that differ in terms of their statistical properties from the *primary* random variables used in microeconomic studies. One important difference between primary and secondary binary variables is that while the former are, in many instances, independently distributed (i.d.) the later are rarely i.d. We show how popular rules for constructing binary states determine the degree and nature of the dependence in those states. When using constructed binary variables as regressands a common mistake is to ignore the dependence by using a probit model. We present an alternative non-parametric method that allows for dependence and apply that method to the issue of using the yield spread to predict recessions.

Key Words: Business cycle; binary variable, Markov chain, probit model, yield curve

JEL Code C22, C53, E32, E37

---

\*We would like to thank a referee and an Associate Editor for constructive comments on an earlier version of the paper.

<sup>†</sup>\*University of Melbourne

<sup>‡</sup>Queensland University of Technology and University of New South Wales. Research supported by ESRC Grant 000 23-0244.

# 1 Introduction

Macroeconometric and financial econometric research often feature binary random variables. We will designate such a random variable as  $S_t$ , and assume that it takes the values of unity and zero. Such binary random variables arise in a number of ways, although they differ in their origin. Because of this it is useful to distinguish between binary random variables that are *primary* and those that are *secondary* or *constructed*. In the first set one would include most of those that arise in micro-econometrics. If a time series is involved there will generally be a panel of data on whether an individual makes a particular decision. In these cases the binary variable is often thought of as deriving from an underlying *continuous latent variable* (as in the Probit model). Also in this set would be cases where a continuous random variable - on which there are realizations - depends upon a latent binary random variable. The clearest example of the latter would be Markov Switching (MS) models - Hamilton (1989). In contrast to those cases, this paper is concerned with secondary binary random variables which are constructed from the realizations of a continuous random variable (or variables)  $y_t$ . This case does not seem to have been studied much, a notable exception being Kedem (1980). However, as we will try to illustrate, quite a few interesting econometric issues arise when such variables are used in empirical work.

Are there many examples of constructed binary time series? There seem to be quite a few, among which we can mention the following.

1. Cycles in economic activity. Here a series  $y_t$  is chosen to represent economic activity and a cycle in it involves expansions,  $S_t = 1$ , and contractions,  $S_t = 0$ . In the event that the series  $y_t$  represents the level of economic activity then it is the *business cycle* that is being isolated. If a permanent component is taken away from  $y_t$  we are investigating the *growth cycle*. In the case of the NBER's dating of the business cycle the variable used for  $y_t$  is the equivalent of the log of GDP - see *The NBER's Recession Dating Procedure at <http://www.nber.org/cycles/recessions.html>*.
2. Bull and bear markets. The underlying variable here will be some asset price e.g. the Dow-Jones or the S&P500 and similar sets of rules as in dating business cycles can be used to perform the segmentation of history into periods of bull and bear markets.

3. Financial crises. Here a unity indicates that a crisis is occurring while a zero indicates that this is not a crisis period -see Eichengreen et al (1995) and Kaminsky and Reinhart (1999) and Bordo et al (2001). The latter state (p 55) that “We construct the familiar index of exchange market pressure (calculated as a weighted average of exchange rate change, short-term interest rate change, and reserve change...). A crisis is said to occur when this index exceeds a critical threshold”.
4. IPO markets are often classified as hot ( $S_t = 1$ ) and cold ( $S_t = 0$ ) depending upon either the volume of new offers or the excess returns earned on them - see Ibbotson and Jaffe (1975) and Brailsford et al (2001).
5. Commodity and real estate markets are often classified as booms and slumps depending upon movements in the underlying prices e.g. Cashin et al. (2002).

One could continue on in this vein but as the examples above indicate, there are many situations in which binary random variables are constructed from some observed continuous random variable. The prevalence of them raises the issue of why this is such a popular strategy. We might put forward a number of reasons:

1. The  $S_t$  may be chosen to emphasize some feature in  $y_t$  that is not immediately obvious e.g. in U.S. business cycles expansions are not smooth but generally feature a period of very fast growth - see Sichel (1994) and Harding and Pagan (2002). This has also been observed in bull markets- see Pagan and Sussonov (2003).
2. Meaningful to decision makers. Because of the well documented phenomenon of loss aversion it is probably not surprising that decision makers are very sensitive to whether there has been a decline (turning point) in series such as GDP and the S&P. Reactions to such an event from the electorate or clients are often very strong and this has led to great interest in being able to predict these events and to examine their causes. This motivates why one might wish to determine the DGP of the  $S_t$  given a known DGP for  $y_t$ .

3. Often the  $S_t$  are objects of interest. An example would be if one wanted to ask whether cycles are synchronized across sectors or countries. Because business cycle dating agencies like the NBER utilize many series in determining the month that a turning point occurred, it is more convenient to examine the coherence of the two cycles, as measured by their representative  $S_t$ , than to try to find correlations with the underlying series that they might have been derived from, since the latter may not even be known to an outside observer.
4. Sometimes there may be large short lived movements in  $\Delta y_t$  that can affect statistics based upon  $\Delta y_t$  but which have little effect upon the constructed  $S_t$  e.g. the stock market crash of October 1987 and the decline in output during the Great Depression. In this instance one might wish to obtain a more robust measure of some feature using the  $S_t$  rather than the  $\Delta y_t$ .
5. There are also many situations in which binary variables are used as inputs into a measure of fit. By far the most common examples are those assessing predictive success. Thus Pesaran and Timmerman (1992) have a sign test of predictive accuracy which has been used to compare output gap estimates by Camba-Mendez and Rodriguez-Palenzuela (2003).

In the next section we will discuss ways of constructing the  $S_t$  from the  $y_t$ . We will distinguish two classes of methods for doing this that are referred to as *turning point* and *termination* rules and give some illustrations of these in the contexts distinguished above. The nature of these rules turns out to be very important in the determination of the DGP of  $S_t$ , and the latter always needs to be carefully derived from that of  $y_t$ , since it is unlikely that the DGPs of  $y_t$  and  $S_t$  will be the same. In particular, it is rare for a constructed  $S_t$  to be *i.d.*, as is typically assumed in micro-econometrics. Moreover, one needs to be wary of using the  $S_t$  as a regressor since  $S_t$  will certainly be a function of  $y_t$  and perhaps  $y_{t+j}$ . In most instances one needs to treat  $S_t$  as an endogenous variable.

As should be expected the DGP of  $S_t$  will be determined by the interaction of the dating rule and the DGP for  $y_t$ . Section 3 provides some illustrations of this, using a combination of theoretical analysis and an examination of some of the actual  $S_t$  which are used in work connected with business cycles, stock market cycles and financial crises. A failure to make an allowance for the fact

that  $S_t$  is not *i.d.* is therefore a potential problem with many existing studies using these variables. Even when there is little serial correlation in the  $S_t$  one also needs to be cautious when using  $S_t$  as regressors as it will generally be a function of  $y_t$  and so, when  $y_t$  is endogenous, so will  $S_t$  be. Section 4 then uses the principles established in section 3 to look at the econometric issues that arise when the  $S_t$  are used in regression models. Two applications in Section 5 close the paper. One concerns the prediction of business cycle states with the yield spread and the other is the use of such states in "Qual-VARs"-Deuker (2005).

## 2 Constructing the States

### 2.1 Rules for Forming the States

The variable  $S_t$  is found by *segmenting* a period of time  $t = 1, \dots, T$  into a history of binary outcomes using information on some underlying continuous random variable. This segmentation requires a *rule* and, depending on what one is studying, this could be classified as either a *turning point* or a *termination* rule. A turning point rule performs the segmentation based upon the location of local maxima and minima in the series  $y_t$ . A termination rule is one which prescribes an event which would cause a change in the value of the state  $S_t$ . In turn termination rules could either be *non-parametric* or derive from a *parametric* model of  $y_t$ .

To illustrate these suppose we consider the  $S_t$  that define a cycle. Perhaps the simplest definition is what might be termed the *calculus rule*. This says that a peak in a series occurs at time  $t$  if  $\Delta y_t > 0$  and  $\Delta y_{t+1} < 0$ . The reason for the name is the result in calculus that identifies a maximum with a change in sign of the first derivative from being positive to negative. A trough (or local minimum) can be found using the outcomes  $\Delta y_t < 0$  and  $\Delta y_{t+1} > 0$ . The states  $S_t$  are simply defined in this case as  $S_t = 1(\Delta y_t > 0)$ , so that  $S_t$  depends only on contemporaneous information. This rule has been popular when  $y_t$  is yearly data, see Cashin and McDermott(2002 ) and Neftci (1984).

When data occurs at (say) the quarterly or monthly frequency one needs to recognize that common usage of a word like "recession" would identify it with a *sustained* decline in the *level* of economic activity i.e. something that lasts for several periods. If one applied the calculus rule there would be too many turning points since the growth rate might regularly switch sign

between one period and the next. Visualizing a peak in a series leads one to the idea that a local peak in  $y_t$  occurs at time  $t$  if  $y_t$  exceeds values  $y_s$  for  $t-k < s < t$  and  $t+k > s > t$ , where  $k$  delineates some symmetric window in time around  $t$ . One can define a trough in a similar way. By making  $k$  large enough we also capture the idea that the level of activity has declined (or increased) in a sustained way. Of course we need to limit the window in time over which this test is applied when performing the test. In this instance  $S_t$  depends upon  $y_{t\pm j}, j = 0, \dots, k$  and so future values of  $y_t$  are needed to determine the value of  $S_t$  i.e. to know whether a turning point occurred at time  $t$  we need to know the future behavior of  $y_t$ . Exactly what amount of future information is required depends on the window width.

A turning point rule based on a non-zero window is the basis of the NBER business cycle dating procedures summarized in the Bry and Boschan (1971) dating algorithm. In that program, designed for the analysis of monthly data,  $k = 5$ . However, because much analysis is conducted with quarterly data, the analogue would seem to be  $k = 2$ . We will refer to this latter rule as the BBQ rule. It is an automated dating rule and therefore differs from the NBER Dating Committee's since the latter utilizes a number of series and exercises judgement. But the correspondence in dates produced by the automated procedure (BBQ) and the NBER choices is close and the situation is reminiscent of the use of an interest rate rule to describe the Fed's setting of the Federal Funds rate. It captures the essence of decisions without the fine detail.

As seen above The calculus rule can also be formulated as a termination rule by expressing it as

$$S_t = 1(\Delta y_t > 0) \tag{1}$$

where  $I(A)$  is the indicator function having value one if the event  $A$  is true and zero otherwise. There is no dependence here on  $S_{t-1}$ . A termination rule that does have such dependence is the "two quarters rule" that often appears in the financial press and which can be summarized as

$$\begin{aligned} S_t &= 1 \text{ if } (\Delta y_{t+1} > 0, \Delta y_{t+2} > 0 | S_{t-1} = 0). \\ S_t &= 0 \text{ if } 1(\Delta y_{t+1} < 0, \Delta y_{t+2} < 0 | S_{t-1} = 1) \\ S_t &= S_{t-1} \text{ otherwise.} \end{aligned} \tag{2}$$

All of the rules above are non-parametric in the sense that they simply look for patterns in the data without making any assumptions about the DGP of  $y_t$ . Parametric (model-based) termination rules proceed by working with a parametric model of  $\Delta y_t$ . Perhaps the best known of these arises by assuming that  $\Delta y_t$  is a function of a latent binary variable  $\xi_t$  that follows a Markov chain and to then construct a series of binary states using the *MS rule*  $\zeta_t = 1[\Pr(\xi_t = 1|F_t) - .5]$ , where  $F_t$  is a set containing either the past history of the observed random variable  $\Delta y_t$  or perhaps the complete sample of observations - see Hamilton (1989). Of course one could use other parametric models of  $\Delta y_t$  to produce  $\zeta_t$  e.g. a SETAR model and the threshold for  $pr(\xi_t = 1|F_t)$  would be different from 0.5. In all of these cases a classification into binary outcomes is produced which is based on whether movements in some function of the  $\Delta y_t$  (and its lags) exceeds a threshold, and the magnitude of the movements involves the parameters of the model.

Each type of rule generates binary random variables but they will not be the same. For this reason we will use the symbol  $S_t$  to designate those that come from either a turning point or non-parametric termination rule and reserve  $\zeta_t$  for those that come from a parametric termination rule. Applied to the same data series  $y_t$  the states  $\zeta_t$  and the states  $S_t$  are conceptually distinct but, in practice, they are often quite close. Thus the  $\zeta_t$  states estimated in Hamilton (1989) with his MS-based termination rule were close to the  $S_t$  coming from using the NBER type rules. Harding and Pagan (2003a) looked at the way in which they differed by using some approximations for getting the  $\zeta_t$  from the history of  $\Delta y_t$ . From that analysis it was clear that the MS rule used a broader information set than the NBER-type turning point rule (in the sense that the latter uses  $\{\Delta y_{t\pm j}\}_{j\leq 2}$  whereas the former uses  $\{\Delta y_{t\pm j}\}_{j\leq T}$ , with downweighting as  $j$  rises). Notice that the latent states in the MS model  $\xi_t$  are *not the same* as the  $\zeta_t$  and so  $\Pr(\xi_t = 1) \neq \Pr(\zeta_t = 1)$  - failure to recognize this is a common error in many studies that use parametric dating rules. Indeed, it is often asserted that the duration of time spent in the state  $\zeta_t$  (or  $S_t$ ) can be determined from the transition probabilities associated with  $\xi_t$ . It is easy to see that this is incorrect since the former will depend on all the parameters of the MS process, including the mean values of  $\Delta y_t$  in each of the regimes, whereas the transition probabilities for  $\xi_t$  do not depend on the mean values. We will focus upon  $S_t$  type measures in this paper but everything said about these holds for the  $\zeta_t$  type measures.

Another important feature of constructed states is that extra censoring rules concerning the minimum or maximum time that can be spent in a

particular state are often applied. Thus in the case of the business cycle dating by the NBER, recessions and expansions must be five months long and a complete cycle must last for 15 months. In quarterly terms these are best interpreted as requiring two quarters as minimum phase lengths and 5 quarters for a complete cycle. We will illustrate how these impact upon the DGP for  $S_t$  later using the quarterly versions.

## 2.2 Examples of State Generation

As we have already demonstrated the measurement of business cycles involves the determination of a set of observations on states that represent expansions and contractions. Bull and bear markets are also popular. Pagan and Sussonov (2003) and Bordo and Wheelock (2006) provide a set of rules for locating turning points in the equivalent of the nominal and real S&P500 respectively, while Lunde and Timmermann(2004) use a non-parametric termination rule. There is also a literature which uses parametric termination rules e.g. the MS model in Maheu and McCurdy(2000).

There have been quite a few adaptations of the approach to stock markets to study booms and slumps in commodity markets - a comprehensive account is in Cashin, McDermott and Scott (2002). Hot and cold markets for IPO's are found by a non-parametric termination rule in Ibbotson and Jaffee (1975) where a hot market is determined by whether excess returns and their changes for two periods exceed the median values while Brailsford et al (2001) used a parametric termination rule ( an MS model).

Non-parametric termination rules to construct the indicators of financial crises most often involve a consideration of the size of the movements in a combination of a number of series. Thus Eichengreen et al. (1995) define a crisis as occurring whether a weighted average of changes in exchange rates, reserves and interest rates exceeds some threshold value. Parametric termination rules have also been applied, mainly based on an MS model e.g. Abiad (2003).

## 3 The DGP of The Binary States

As mentioned in the introduction the DGP of constructed binary states is not one that the investigator is free to prescribe. It is determined by the interaction of the DGP of the variable they are constructed from and the type

of rule used for mapping the observable variable into the binary states. As we will note the DGP is generally a high-order Markov process. Nevertheless, it is useful for understanding the DGP to think of approximating the higher order process with a first order one. This is not an unfamiliar process. A high order autoregression in a continuous random variable can always be approximated by a first order AR and, if one wished (say) to measure the degree of persistence in the process, this approximation often gives a very good indication of that quantity. We follow this approximation strategy in the next sub-section.

### 3.1 Serial Correlation in the States

If the states evolve as a first-order Markov process then Hamilton (1994 p684) shows that the following identity holds:

$$S_t = p_{01} + (1 - p_{01} - p_{10})S_{t-1} + \eta_t, \quad (3)$$

where  $\eta_t$  is discrete and conditionally heteroskedastic since it depends upon  $S_{t-1}$  and

$$p_{jk} = \Pr(S_{t+1} = k | S_t = j) \quad (4)$$

The determinants of  $p_{jk}$  will depend upon the nature of the DGP for  $y_t$  and the type of rule employed to construct  $S_t$ . To illustrate this we suppose that

$$\Delta y_t = \mu + \sigma e_t \quad (5)$$

where  $e_t$  is *i.i.d.*(0,  $\sigma^2$ ). Now if the calculus rule is employed i.e.  $S_t = 1(\Delta y_t > 0)$ ,

$$\begin{aligned} p_{10} &= \Pr(S_{t+1} = 0 | S_t = 1) \\ &= \Pr(\Delta y_{t+1} < 0 | \Delta y_t > 0) \\ &= \Pr(\Delta y_{t+1} < 0) = \psi \end{aligned}$$

due to independence of  $\Delta y_t$ . In the same way  $p_{01} = 1 - \psi$  and, from (3),

$$S_t = 1 - \psi + (0 \times S_{t-1}) + \eta_t, \quad (6)$$

showing that there is no serial correlation in the states  $S_t$ .

Now, what happens if one relaxes the assumption that  $y_t$  follows a random walk with drift? Using the calculus rule, combined with  $\Delta y_t$  being a

mean-zero stationary Gaussian process, Kede(1980, p34) sets out the relation between the autocorrelations of the  $\Delta y_t$  and  $S(t)$  processes. Letting  $\rho_{\Delta y}(k) = \text{corr}(\Delta y_t, \Delta y_{t-k})$ , and  $\rho_S(k) = \text{corr}(S_t, S_{t-k})$ , he determines that

$$\rho_S(k) = \frac{2}{\pi} \arcsin(\rho_{\Delta y}(k)). \quad (7)$$

Thus, given an estimate of  $\rho_{\Delta y}(k)$ , we can immediately find an estimate of  $\rho_S(k)$  and *vice versa*, making it clear that an AR process for  $\Delta y_t$  will result in a much more complex DGP for  $S_t$  than an  $AR(1)$ .

So the autocovariances in  $S_t$  depend upon whether there is serial correlation in  $\Delta y_t$ . But, even if there is no serial correlation in  $\Delta y_t$ , the dating rule itself can induce it into  $S_t$ . The analysis will be done assuming that  $y_t$  follows (5) and the ‘‘two quarters rule’’ for dating phase shifts. The complications in working with this rule come from the fact that the conditioning event  $S_{t-1} = 1$  will place some restrictions upon the signs of past sample paths for  $\{\Delta y_t\}$  that are associated with an expansion terminating sequence defining the move from  $S_{t-1} = 1$  to  $S_t = 0$ . For example the sequence

$$\{\Delta y_{t+1}, \Delta y_t, \Delta y_{t-1}, \Delta y_{t-2}, \dots\} = \{-, -, -, +, \dots\} \quad (8)$$

would be incompatible with  $S_{t-1} = 1$ , since the negative growth at  $t - 1$  would match with the negative growth at  $t$ , and so the expansion would have been terminated at  $t - 1$ . The appendix shows that

$$p_{10} = \frac{\psi^2}{(1 + \psi)}, \quad p_{01} = \frac{(1 - \psi)^2}{2 - \psi} \quad (9)$$

$$p_{11} = \frac{1 + \psi - \psi^2}{(1 + \psi)}, \quad p_{00} = \frac{1 + \psi - \psi^2}{2 - \psi}. \quad (10)$$

Hence, using (3), we will have

$$S_t = \frac{(1 - \psi)^2}{2 - \psi} + \left[1 - \frac{(1 - \psi)^2}{2 - \psi} - \frac{\psi^2}{(1 + \psi)}\right] S_{t-1} + \eta_t \quad (11)$$

To get some feel for the magnitude of the coefficients in this relation assume that  $\Delta y_t$  is  $N(\mu, \sigma^2)$ , so that  $\psi = \Phi(-\frac{\mu}{\sigma})$ , where  $\Phi(u)$  is the cumulative standard normal distribution function. Using sample estimates of  $\mu$  and  $\sigma^2$  for US GDP over the period 1959/1-1997/2 (the same sample as used in the

application by Estrella and Mishkin (1998)) gives  $\psi = .21$ . Inserting this into (11) produces

$$S_t = .35 + .62S_{t-1} + \eta_t, \quad (12)$$

showing that there is substantial serial correlation in the states. Fitting an AR(1) to the quarterly “NBER business cycle states” (found from their web page) over the same period yields

$$S_t = .29 + .67S_{t-1} + \eta_t, \quad (13)$$

which shows that the predictions about the nature of the business cycle states identified by the NBER, and those using the “two quarters rule” are quite good.

These results also continue to hold for the  $S_t$  found from the monthly S&P500 using the turning point definitions in Pagan and Sussonov (2003). The regression over 1854/6-1997/12 gives

$$S_t = .07 + .89S_{t-1} + \eta_t \quad (14)$$

(8.1) (80.5)

So it is very likely that there will be serial correlation in the states  $S_t$ . This is important since it means that secondary (constructed) states cannot be treated as if they were primary states. In particular, it will not be correct to assume that they are realizations from an *i.d.* process, as has been done in many applications with them. Examples of the latter in the time series literature would be the market timing test of Pesaran and Timmermann (1992) and its close relative, Pearson’s test of independence in a contingency table (see Artis et al (1997)). t ratios underlying these tests are effectively constructed under the *i.i.d.* assumption. In Pesaran and Timmermann’s context this may be a valid assumption, since the  $y_t$  are forecast errors and the  $S_t = 1(y_t > 0)$ . But others have applied it to  $y_t$  that are possibly serially correlated e.g. in Camba-Mendez and Rodriguez-Palenzuela (2003) the  $y_t$  are the revision errors in output gaps and there is no reason to think that these would be serially uncorrelated. In all instances an adjustment needs to be made for the serial correlation in the  $S_t$ . As seen in Harding and Pagan(2006), the requisite adjustment to t-statistics of the Artis et al (1997) test of synchronization of cycles can be very large indeed.

### 3.2 Effects of Censoring Rules on the DGP of the States

Rules that involve restrictions on the duration of time spent in a phase, such as the minimum two-quarter restriction for recessions and expansions used by the “two quarters” and BBQ rules, also place strong restrictions on the nature of the DGP for  $S_t$ . To examine this in more detail, let  $\Delta y_t - \mu$  be a mean-zero covariance stationary process. Then, under most dating rules, the  $S_t$  are generated as nonlinear functions of  $\Delta y_t$  and  $\Delta y_{t+1}$ , and thus  $S_t$  is covariance stationary.

Now let us look at the case where the calculus rule is adopted but some censoring of binary variables is employed to impose phase-length restrictions. We have already mentioned that in these circumstances, and without phase restrictions, Kedem has shown that one can represent  $S_t$  as a  $K$ 'th order Markov process, where  $K$  may need to be infinite. A first order Markov process can always be written as

$$S_t = \mu_0 + \phi_1 S_{t-1} + \varepsilon_t \quad (15)$$

while a second order one has the form

$$S_t = \mu_0 + \phi_1 S_{t-1} + \phi_2 S_{t-2} + \psi_1 S_{t-1} S_{t-2} + \varepsilon_t. \quad (16)$$

Higher order processes consist of the linear form plus all the interaction terms. Notice that terms like  $S_{t-1}^k$  are just  $S_{t-1}$  so that we only need to consider interaction terms using  $k = 1$ .

Using Kedem's result, the binary process  $S_t$  can be represented as

$$S_t = \mu_0 + \phi_1 S_{t-1} + \phi_2 S_{t-2} + f(S_{t-1}, S_{t-2}, \tilde{S}_{t-3}) + \varepsilon_t, \quad (17)$$

where  $\tilde{S}_t = \{S_{t-i}\}_{i=0}^{\infty}$  and the notation  $f(\cdot)$  means the sum of all interaction terms of  $S_{t-1}$  and  $S_{t-2}$  with  $\tilde{S}_{t-3}$  e.g.  $S_{t-1} S_{t-2} S_{t-3}$ , as well as elementary and interaction terms formed from the elements of  $\tilde{S}_{t-3}$  alone e.g.  $S_{t-3}, S_{t-3} S_{t-4}$ . Then

$$E(S_t | \tilde{S}_{t-1}) = \mu_0 + \phi_1 S_{t-1} + \phi_2 S_{t-2} + f(S_{t-1}, S_{t-2}, \tilde{S}_{t-3}) \quad (18)$$

Now the restriction that a phase must last two quarters implies that  $\Pr(S_t = 1 | S_{t-1} = 0, S_{t-2} = 1, \tilde{S}_{t-3}) = 0$ . Since the probability equals  $E(S_t | S_{t-1} = 0, S_{t-2} = 1, \tilde{S}_{t-3})$  we see that

$$\mu_0 + \phi_2 + f(0, 1, \tilde{S}_{t-3}) = 0 \quad (19)$$

and this can only occur for arbitrary  $\tilde{S}_{t-3}$  if  $f(0, 1, \tilde{S}_{t-3}) = 0$ . Hence  $S_t$  must be second order Markov with  $\mu_0 = -\phi_2$ . Since it is also the case that  $\Pr(S_t = 0 | S_{t-1} = 1, S_{t-2} = 0, \tilde{S}_{t-3}) = 0$  we find that  $\mu_0 + \phi_1 = 1$ . Thus the presence of censoring has influenced both the order of the Markov Chain that is the DGP of the states and has also imposed restrictions upon the coefficients of the linear (in parameters) representation that the chain can be given. To check this prediction out we fitted a second order Markov Chain to the business cycle states found by applying the BBQ rule ( with minimum phase restriction of two periods imposed) to US quarterly GDP over 1947/1-2002/2. The resulting parameter estimates clearly satisfy the predicted relations between the coefficients (all coefficients were highly significant):

$$S_t = .45 + .55S_{t-1} - .45S_{t-2} + .40S_{t-1}S_{t-2} + \eta_t. \quad (20)$$

If we had dropped the interactive term then the regression would be

$$S_t = .38 + .78S_{t-1} - .22S_{t-2} + \eta_t, \quad (21)$$

with heteroskedastic robust t ratios for the AR parameters of 13.2 and 3.8 respectively.

The same type of effects can be seen in other series, even when there is no explicit censoring, but rather is due to the fact that a minimum duration to phases might occur within the sample. For example fitting a second order Markov process to the data for financial crises in the United Kingdom over 1883 to 1998 from Bordo et al (2001) we get

$$S_t = .07 + .05S_{t-1} + .05S_{t-2} + .49S_{t-1}S_{t-2} \quad (22)$$

The  $S_{t-1}$  and  $S_{t-2}$  terms are not significant but the interaction term is, and after dropping the insignificant terms we get

$$S_t = .08 + .59S_{t-1}S_{t-2}, \quad (23)$$

with the t ratio on the interaction term being 3.54. Thus the type of serial correlation in the states can be quite complex.

### 3.3 Testing the Order of a Markov Chain

If we assume that  $S_t$  is generated by a  $k'$ th order Markov Chain it can be given the linear form

$$S_t = X_t^k \beta_k + \eta_t \quad (24)$$

where the  $X_t^k$  are generated from a recursion as follows

$$\begin{aligned} X_t^0 &= 1 \\ X_t^j &= [ X_t^{j-1} \quad X_t^{j-1} S_{t-j} ]. \end{aligned}$$

Thus we would have

$$\begin{aligned} X_t^1 &= [ 1 \quad S_{t-1} ] \\ X_t^2 &= [ 1 \quad S_{t-1} \quad S_{t-2} \quad S_{t-1}S_{t-2} ] \\ X_t^3 &= [ 1 \quad S_{t-1} \quad S_{t-2} \quad S_{t-1}S_{t-2} \quad S_{t-3} \quad S_{t-1}S_{t-3} \quad S_{t-2}S_{t-3} \quad S_{t-1}S_{t-2}S_{t-3} ] \end{aligned}$$

Now the number of parameters in a Markov chain grows as  $2^k$ . Consequently, high order processes would be hard to estimate unless there are large numbers of observations. It is also the case that, with relatively small sample sizes, the matrix  $X_t^k$  can be singular, since the summation of some of the terms may be identical. Thus, in testing the order of the business cycle states using the NBER states in Estrella and Mishkin's application one cannot estimate a third order Markov chain as  $S_{t-1}S_{t-2}S_{t-3}$  and  $S_{t-1}S_{t-3}$  are perfectly correlated.

Various suggestions have been made to reduce the number of parameters to be estimated in the Markov process. Raftery (1985) suggests a parameterization that reduces the number of parameters from  $2^k$  to  $2^{k-1}$ . It is also possible to get such a reduction by imposing a minimum phase restriction and it appears that this is an alternative interpretation of what Raftery's restrictions do ( he does not provide an interpretation). The restrictions matrix has the form,

$$R_k \beta = r_k \tag{25}$$

and can be built up via recursion as follows. Starting with the restrictions on a second order system of

$$R_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}, \quad r_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{26}$$

those on higher order processes are generated via the following recursion:

$$R_j = \begin{bmatrix} R_{j-1} & 0 \\ R_{j-1} & R_{j-1} \end{bmatrix}, \quad r_j = \begin{bmatrix} r_{j-1} \\ r_{j-1} \end{bmatrix} \tag{27}$$

Since there are  $2^k$  parameters in the unrestricted  $k^{th}$  order Markov chain there will be  $2^{k-1}$  parameters in the restricted version and thus the latter

may be possible to estimate when the former is not. By multiplying  $R_j$  and  $r_j$  by the matrix  $T_j$  defined by

$$T_j = \begin{bmatrix} I & 0 \\ -I & I \end{bmatrix}, \quad (28)$$

we obtain a simpler expression for the restrictions of the form:

$$R_j^* \beta_j = r_j^*, \quad (29)$$

where

$$R_j^* = \begin{bmatrix} R_{j-1} & 0 \\ 0 & R_{j-1} \end{bmatrix}, r_j^* = \begin{bmatrix} r_{j-1} \\ 0 \end{bmatrix}. \quad (30)$$

When estimating the model with restrictions imposed it is useful to re-write the model as

$$\beta_j = C_j \alpha_j + d_j, \quad (31)$$

where  $R_j C_j = 0$  and  $R_j d_j = r_j$ . Then  $C_j$  can be obtained recursively as,

$$C_j = \begin{bmatrix} C_{j-1} & 0 \\ 0 & C_{j-1} \end{bmatrix}. \quad (32)$$

Since  $T_j R_j d_j = T_j r_j$  it follows that

$$\begin{bmatrix} R_{j-1} & 0 \\ 0 & R_{j-1} \end{bmatrix} \begin{bmatrix} d_{j-1} \\ d^* \end{bmatrix} = \begin{bmatrix} r_{j-1} \\ 0 \end{bmatrix} \quad (33)$$

This must imply that  $d^* = 0$  and so  $d_j$  can be generated recursively as

$$d_j = \begin{bmatrix} d_{j-1} \\ 0 \end{bmatrix}. \quad (34)$$

Thus,  $C_2$ ,  $d_2$  and the recursion relations (32) and (34) are all that are required to construct the restriction matrices.

For later use we wish to determine the order of the Markov chain in the business cycle states used by Estrella and Mishkin. We initially started with a fourth order Markov chain as the general model, but found that one could not estimate this model, even with the phase length restrictions imposed. Thus the maximum order of the chain needed to be reduced to three, in which case it has the form

$$\begin{aligned}
S_t = & \beta_1 + \beta_2 S_{t-1} + \beta_3 S_{t-2} + \beta_4 S_{t-1} S_{t-2} + \beta_5 S_{t-3} \\
& + \beta_6 S_{t-1} S_{t-3} + \beta_7 S_{t-2} S_{t-3} + \beta_8 S_{t-1} S_{t-2} S_{t-3}.
\end{aligned} \tag{35}$$

This model was estimated (using Ordinary Least Squares) after imposing the two quarter minimum phase restrictions i.e.

$$\begin{aligned}
\beta_1 + \beta_2 &= 1 \\
\beta_1 + \beta_3 &= 0 \\
\beta_1 + \beta_2 + \beta_5 + \beta_6 &= 1 \\
\beta_1 + \beta_3 + \beta_5 + \beta_7 &= 0.
\end{aligned}$$

Imposing these restrictions on the third order Markov process the residual sum of squares was 9.228. We then estimated a restricted second order process and, as it had virtually the same sum of squares, it was easy to accept a second order restricted process. Proceeding to the next order in the sequence it was easy to reject the hypothesis of a first order process since the sum of squares from a regression of  $S_t$  on a constant and  $S_{t-1}$  was 9.99. One question that might arise however is whether the dependence in the process for  $S_t$  might be somewhere between a second and third order Markov chain. We therefore asked if the second order process might be augmented by the variables  $S_{t-3}$  and  $S_{t-2}S_{t-3}$ . The heteroskedasticity adjusted Wald test for this is 6.26, with a  $p$  value of .044. Given the fact that this is only an asymptotic test it seems reasonable to conclude that a second order process is a reasonable approximation for this data set.

## 4 Applications Using the States

So far we have demonstrated how the DGP of the constructed states needs to be treated quite carefully. In particular, because the states are constructed variables, they cannot be treated in the same way as they would be in microeconomic work, where information is directly available (say) on whether a person is unemployed or not. The states have also often been used in regressions, either as regressors or as the dependent variable, and we therefore need to look at the implications of the results established in the previous sections for these uses. Finally in this section we look at how one can use the

$S_t$  to produce forecasts. It is often said that the advantage of a parametric model is that one can easily produce forecasts from it since the available information can be conditioned upon. We aim to show that it is possible to produce forecasts of the constructed binary state outcomes using only information available to a forecaster. This is obviously important if one is trying to forecast a recession but also if one is trying to assess the chances of a financial crisis using "early warning" indicators.

## 4.1 Constructed Binary Variables as Regressors

There is an emerging tendency to utilize  $S_t$  as regressors. In these applications they are included in a regression such as

$$y_t = a + bx_t + cS_t + dx_tS_t + e_t, \quad (36)$$

where the effect of  $x_t$  upon  $y_t$  may change according to the value of  $S_t$ . Thus one might have either  $y_t$  as output and  $x_t$  an interest rate or  $y_t$  might be inflation and  $x_t$  an output gap. Such possibilities are often mentioned. In particular there have been tests for the asymmetric effects of monetary policy e.g. Cover (1992) but in the past these tests have been done through a definition like  $S_t = 1(w_t > 0)$ , where  $w_t$  might be  $e_t$  or  $\Delta y_t$ . Clearly such tests do not effectively address whether the impact of monetary policy is different in different phases of the cycle, since the resulting  $S_t$  do not delineate the business cycle phases.

To illustrate the complications that are caused by using  $S_t$  as a regressor we assume it has been generated with the BBQ rule. Then we know that

$$S_t = 1(\Delta_2 y_t > 0, \Delta y_t > 0, \Delta_2 y_{t+2} < 0, \Delta y_{t+1} < 0). \quad (37)$$

It is therefore clear that we cannot use  $S_t$  as a regressor, since it is a function of  $e_{t+2}$ ,  $e_{t+1}$  and  $e_t$ . It would however be possible to use  $S_{t-3}$  as an instrument for  $S_t$ .

Even if the  $S_t$  do not appear explicitly as a regressor in equations they are often implicitly there. Thus Dueker(2005) has a Qual-VAR which has the (simplified) form

$$\begin{aligned} y_t &= \alpha_{yy}y_{t-1} + \alpha_{yz}z_{t-1} + \varepsilon_t \\ z_t &= \alpha_{zy}y_{t-1} + \alpha_{zz}z_{t-1} + u_t \\ \zeta_t &= 1(z_t > 0), \zeta_t = S_t \end{aligned}$$

where  $z_t$  is a latent variable and the shocks are normally and independently distributed with a zero expectation. In his application  $S_t$  are the NBER business cycle states and he estimates the parameters of the Qual-VAR using data on  $y_t$  and  $S_t$ . Estimation is Bayesian and done using MCMC methods. The values of  $S_t$  determine what distribution one should draw  $z_t$  from i.e. at time  $t$ , if we observe that  $S_t = 1$ , we draw from the truncated distribution  $z_t|z_t > 0$ . In this sense the  $S_t$  are being treated as if they are exogenous even though we know that they have been constructed from the  $y_t$ . The implications of the model above are of course

$$\begin{aligned} E(\zeta_t|y_{t-1}, z_{t-1}) &= \text{prob}(z_t > 0) \\ &= \Phi(\alpha_{zy}y_{t-1} + \alpha_{zz}z_{t-1}) \end{aligned}$$

so that, adopting a linear approximation,

$$E(\zeta_t|y_{t-1}, z_{t-1}) = ay_{t-1} + bz_{t-1},$$

produces an explanation of  $\zeta_t$  of the form

$$\zeta_t = S_t = ay_{t-1} + bz_{t-1} + \eta_t,$$

where  $\eta_t = \zeta_t - E(\zeta_t|y_{t-1}, z_{t-1})$ . Inverting this equation gives  $z_{t-1} = \frac{1}{b}(S_t - ay_{t-1} - \eta_t)$  so that the model for  $y_t$  becomes

$$y_t = cy_{t-1} + dS_t + \xi_t. \tag{38}$$

Consequently, even though  $S_t$  did not appear explicitly in the original VAR, it was implicitly there because of the presence of the latent variable  $z_{t-1}$ .

Now if  $\alpha_{zy} = 0$  we know from Kadem (1980) that  $\zeta_t$  will be an infinite dimensional Markov Chain whose parameters depend only upon  $\alpha_{zz}$  and so the combination of (38) and the stationary process for  $S_t$  will effectively be the system that is being estimated. Conditioning upon  $S_t$  raised the issue of whether  $S_t = \zeta_t$ . We know that there will be an estimate of  $\zeta_t$  that can be generated from  $\{y_{t-j}\}_{j=0}^{\infty}$  but whether that corresponds to the way that the NBER construct  $S_t$  from  $y_t$  is problematic. There is probably some specification for the  $z_t$  process that will lead to a dating rule using  $y_t$  that will agree with the NBER states but whether it is the one in the Qual-VAR is another matter but it does point to the need to check for specification errors in latent variable part of the Qual-VAR. In any cases it is clear that the

fact that  $S_t$  is being treated as if it is exogenous when it depends on future values of  $y_t$  will lead to inconsistent estimation of  $d$  unless one recognizes this dependence. One way of doing this, which also preserves the NBER dating rule, would be to use a VAR in  $y_t$  and  $S_t$  as an auxiliary model, and to estimate the Qual-VAR by indirect estimation, where simulated data from the Qual-VAR is passed through some dating algorithm that produces  $S_t$  from  $y_t$ .

## 4.2 Constructed Binary Variables as Regressands

Often  $S_t$  are the variables to be explained or predicted and it is desired to test if the regressors have an influence upon  $S_t$ . A simple case of this is testing for synchronization of cycles where  $x_t$  is another cyclical indicator. It is clear that, if one runs a regression of  $S_t$  upon  $x_t$ , and tests the null hypothesis that the coefficient of  $x_t$  is zero in such a regression, then one must take account of the fact that, under the null hypothesis,  $S_t$  has extensive serial correlation and heteroskedasticity and test statistics must be made robust to those features. The necessary adjustments were shown to be large in Harding and Pagan (2006).

Now motivated by the micro-econometric literature it is often felt desirable to test if  $\Pr(S_t = 1)$  is a function of some determinants  $x_t$ . Historically, those working with *constructed states* followed the micro-econometrics approach, in which it is first postulated that  $\Pr(S_t = 1) = F(-x_t'\beta)$ , where  $F(\cdot)$  is a c.d.f., and then a likelihood is established under the assumption that  $S_t$  are *i.i.d.* An alternative is to assume that there is some latent variable process  $y_t^*$  that is a linear function of a single index  $x_t'\beta$  with the format

$$y_t^* = x_t'\beta + u_t^*, \quad (39)$$

where  $u_t^*$  are *i.i.d.* Then a density for  $u_t^*$  is prescribed and  $S_t = 1(y_t^* > 0)$  is taken to be the rule for generating the  $S_t$ . The  $F(\cdot)$  of the first approach is then just the distribution function corresponding to the density function for  $u_t^*$ . Examples of these methodologies are Estrella and Mishkin (1998), Birchenall et al (1999) and Chen et al (2000).

We can see that there are two problems with these approaches. One is that the  $S_t$  are not *i.i.d.* and so one cannot write a likelihood as

$$\prod_{t=1, S_t=1}^T F(-x_t'\beta) \prod_{t=1, S_t=0}^T (1 - F(-x_t'\beta)). \quad (40)$$

Nevertheless, it is this form that has been used in the literature to date. Chen et al and Birchenall et al. are explicit about the fact that  $S_t$  is a constructed variable in time series but then ignore the method of construction when specifying and estimating the models that seek to explain  $S_t$ . If the  $S_t$  were first order Markov the likelihood could be formed as the product of transition probabilities distinguishing the various state shifts that can occur. In theory one might do this for higher order processes, but it would become increasingly complex. Moreover, as we have emphasized before, it seems likely that, in many cases, we would not know the exact order of the Markov process. An alternative is to work with the latent variable model as there is a literature that allows it to be serially correlated. If the dating rule is known it may be possible to find the likelihood using computer simulation methods. We have worked through a simple case in the Appendix where we found that the transition probabilities at time  $t$  would depend on the complete past history of  $x_t$ . But, in general it will be very difficult to derive a likelihood for  $S_t$  conditional upon  $x_t$ , simply because we often know little about which  $y_t$  are used in the construction of the  $S_t$  and the precise dating rules that are used.

Some feasible method of accounting for the nature of  $S_t$  in determining relations with  $x_t$  is needed, even if it is approximate and not fully efficient. We will assume that the states are  $S_t$  and the question to be answered is whether transition probabilities vary with some regressor  $x_t$ . Thus we have

$$\Pr(S_t|\tilde{S}_{t-1}, x_t) = h(\tilde{S}_{t-1}, x_t), \quad (41)$$

where  $\tilde{S}_t$  is some conditioning set whose nature depends upon the order of the Markov chain.

Now we know that, for any finite order Markov chain, if  $g(x_t) = 0$  then we can write

$$\Pr(S_t|\tilde{S}_{t-1}) = E(S_t|\tilde{S}_{t-1}) = \tilde{S}'_{t-1}\delta. \quad (42)$$

Thus, for the second order case,

$$\tilde{S}'_{t-1}\delta = S_{t-1}\delta_1 + S_{t-2}\delta_2 + S_{t-1}S_{t-2}\delta_3. \quad (43)$$

This suggests that we consider a separable version of  $h(\cdot)$ , allowing the transition probability to be written as

$$\Pr(S_t|\tilde{S}_{t-1}, x_t) = \tilde{S}'_{t-1}\delta + g(x_t). \quad (44)$$

This has the regression format

$$S_t = \tilde{S}'_{t-1}\delta + g(x_t) + u_t, \quad (45)$$

where  $E(u_t|\tilde{S}_t, x_t) = 0$ .

We then wish to estimate  $g(x_t)$ . To do this we can use the semi-parametric method of Robinson (1988). This proceeds by taking the expectation of  $S_t$  given  $x_t$ ,

$$E(S_t|x_t) = E(\tilde{S}_{t-1}|x_t)'\delta + g(x_t) \quad (46)$$

and then forming

$$S_t - E(S_t|x_t) = [\tilde{S}_{t-1} - E(\tilde{S}_{t-1}|x_t)]'\delta + u_t. \quad (47)$$

The conditional expectations in (47) can be estimated quite accurately by non-parametric methods as there is only a scalar as the conditioning element. Once these have been found the regression of  $S_t - E(S_t|x_t)$  against  $[\tilde{S}_{t-1} - E(\tilde{S}_{t-1}|x_t)]$  provides an estimate of  $\delta$ . With this estimate  $g(x_t)$  can be extracted by non-parametrically computing  $E[(S_t - \tilde{S}'_{t-1}\hat{\delta})|x_t]$ .

Notice the importance of the separability assumption. If this is incorrect then we would need to estimate  $h(\cdot)$  by non-parametric methods, and then the number of conditioning elements will depend on the dimension of  $\tilde{S}_t$ , which can be quite high. With a low order Markov process this may be a feasible estimation strategy, but not if the order is above second. The motivation for the separability assumption is that it produces a model which nests those in the literature by setting  $\delta = 0$  and making  $g(\cdot)$  the cumulative normal (when a Probit model is adopted). Thus what we are doing here would be simplest possible generalization that retains the same structure as the Probit models in the literature but allows for dependence in the  $S_t$ . We could obviously compare the estimated  $g(\cdot)$  to the cumulative normal.

As an example we consider the influence of the yield spread upon the probability of moving to a recession. Estrella and Mishkin(1998) did this via a Probit model assuming  $S_t$  were *i.d.*. As the  $S_t$  they used has previously been shown to be second order Markov we therefore fit the model

$$S_t = \delta_1 S_{t-1} + \delta_2 S_{t-1} + \delta_3 S_{t-1} S_{t-2} + g(x_t) + \eta_t \quad (48)$$

where  $x_t$  is the yield spread lagged two quarters. The intercept in this relation is absorbed into  $g(x_t)$  since we do not know the function  $g(\cdot)$ .

Fitting this model using Robinson's method produces  $\delta_1 = .507, \delta_2 = -.33, \delta_3 = .28$ . In order to compare results from this equation with the Probit model used by Estrella and Mishkin we need to compute  $E(S_t = 0) = 1 - E(S_t) = 1 - \mu$ . Taking expectations

$$\mu = \delta_1\mu + \delta_2\mu + \delta_3E(S_{t-1}|S_{t-2} = 1) \Pr(S_{t-2} = 1) + \mu_g,$$

where  $\mu_g = E(g(x_t))$ . Now

$$E(S_{t-1}|S_{t-2} = 1) = \delta_1 + \delta_2ES_{t-3} + \delta_3ES_{t-3} + \mu_g \quad (49)$$

$$= \delta_1 + (\delta_2 + \delta_3)\mu + \mu_g \quad (50)$$

and

$$\Pr(S_{t-2} = 1) = E(S_{t-2}) = \mu, \quad (51)$$

so that

$$\mu = \delta_1\mu + \delta_2\mu + \delta_3(\delta_1 + (\delta_2 + \delta_3)\mu + \mu_g)\mu + \mu_g \quad (52)$$

This is a quadratic in  $\mu$ . We take the root that lies between zero and 1. There is no guarantee that there will be a positive value and some small negative values were found for very large values of  $x$ .

Figure 1 plots  $E(S_t = 0)$  and the Probit estimate of  $\Pr(S_t = 0)$  against  $x_t$ . It is clear that there is a substantial difference between the Probit and non-parametric estimates of  $g(\cdot)$  when the yield spread lies between -1% to 1%. The non-parametric estimate suggests that large negative spreads are needed to produce a high probability of a recession.

### 4.3 Forecasting Recessions

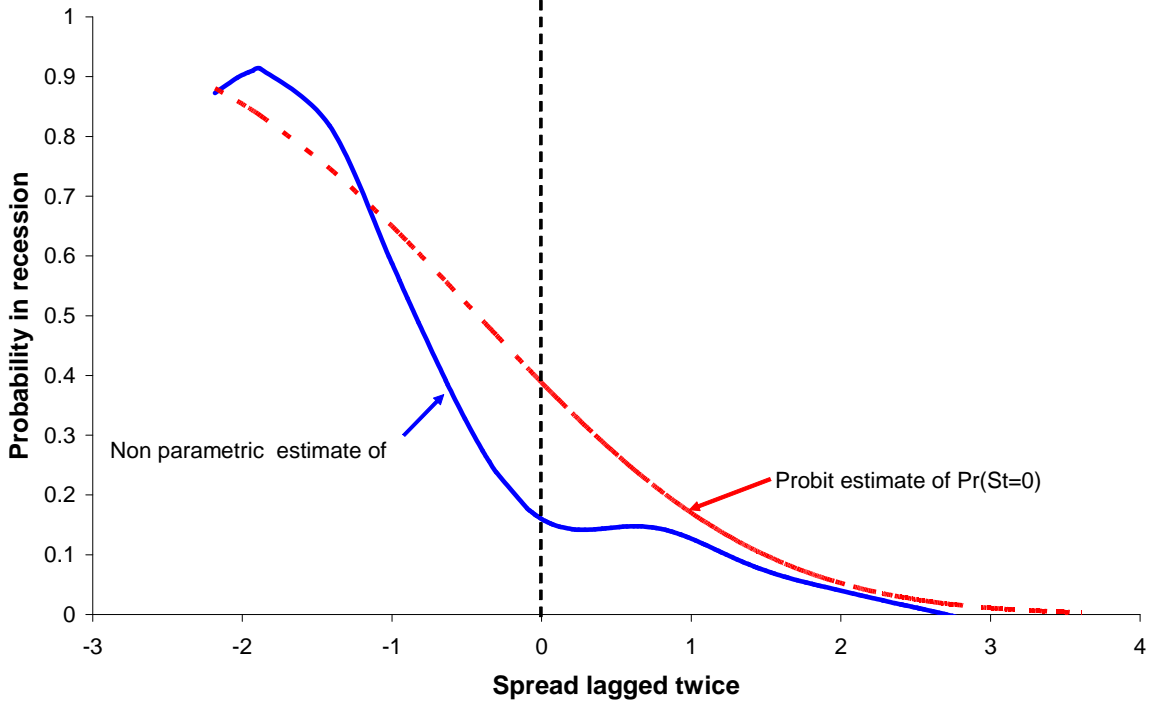
We will look at the issue of how one forecasts the states if one knows that the constructed variable is a function of future values of an endogenous variable. Specifically we study the forecasting of recessions using quarterly data, since this has been dealt with by others and so can provide a comparison. Given our earlier work it will be assumed that the states follow a second order Markov process of the form

$$S_t = a_0 + a_1S_{t-1} + a_2S_{t-2} + a_3S_{t-1}S_{t-2} + g(x_t) + \eta_t, \quad (53)$$

where  $E(\eta_{t-l}|F_t) = 0$  for  $l \leq 2$  and  $F_t$  is the information available for the forecast, which we will take to be the available history on  $S_{t-3}$  and  $x_t$  at time  $t$ . (designated as  $\tilde{S}_t, \tilde{x}_t$  respectively). We seek to forecast  $S_{t+j}, j = 0, 1, \dots$

Beginning with  $t = 0$  the problem with forecasting  $S_t$  directly from this equation is that, although  $x_t$  is known,  $F_t$  does not include  $S_{t-1}$  and  $S_{t-2}$ ,

Figure 1: Probability of a recession, estimates from probit model and non-parametric model



since information is needed about output in time  $t$  and  $t + 1$  to determine their realizations. Taking expectations of (53) conditional on  $[\tilde{S}_{t-3}, \tilde{x}_{t-2}]$  yields

$$E(S_t|F_t) = a_0 + a_1E(S_{t-1}|F_t) + a_2E(S_{t-2}|F_t) + a_3E(S_{t-1}S_{t-2}|F_t) + g(x_t) \quad (54)$$

and any forecast requires us to compute these expectations. Now lagging (53) one period and taking expectations conditional on  $F_t$  yields

$$E(S_{t-1}|F_t) = a_0 + (a_1 + a_3S_{t-3})E(S_{t-2}|F_t) + a_2S_{t-3} + g(x_{t-1}) \quad (55)$$

Next, lag (53) two periods and take expectations conditional on  $F_t$  to obtain

$$E(S_{t-2}|F_t) = a_0 + a_1S_{t-3} + a_2S_{t-4} + a_3S_{t-3}S_{t-4} + g(x_{t-2}) \quad (56)$$

Clearly we can construct an estimate of  $E(S_{t-2}|F_t)$  from (56) and we will call that  $\hat{S}_{t-2}$ . This can then be substituted into (55) to get  $\hat{S}_{t-1}$  so that our forecast of  $S_t$  will be

$$\begin{aligned} E(S_t|F_t) &= \hat{S}_t = a_0 + a_1\hat{S}_{t-1} + a_2\hat{S}_{t-2} \\ &\quad + a_3E(S_{t-1}S_{t-2}|F_t) + g(x_t) \end{aligned} \quad (57)$$

This leaves  $E(S_{t-1}S_{t-2}|F_t)$  to be determined. It is not equal to  $\hat{S}_{t-1}\hat{S}_{t-2}$  since

$$E(S_{t-1}S_{t-2}|F_t) = E(S_{t-1}|S_{t-2} = 1, F_t) \Pr(S_{t-2} = 1|F_t) \quad (58)$$

so that

$$E(S_{t-1}S_{t-2}|F_t) = E(S_{t-1}|S_{t-2} = 1, F_t) \hat{S}_{t-2}.$$

It also follows that

$$E(S_{t-1}|S_{t-2} = 1, F_t) = a_0 + a_1 + (a_2 + a_3)S_{t-3} + g(x_{t-1}). \quad (59)$$

The above equations establish initial conditions for the forecasts and, for  $j > 0$ , it's clear that we can recursively generate these as

$$\begin{aligned} \hat{S}_{t+j} &= a_0 + a_1\hat{S}_{t+j-1} + a_2\hat{S}_{t+j-2} + \\ a_3E(S_{t+j-1}|S_{t+j-2} &= 1, F_t)\hat{S}_{t+j-2} + E(g(x_{t+j})|\tilde{x}_t) \end{aligned} \quad (60)$$

where

$$\begin{aligned} E(S_{t+j-1}|S_{t+j-2} &= 1, F_t) = a_0 + a_1 + (a_2 + a_3)\hat{S}_{t+j-3} \\ &\quad + E(g(x_{t+j-1})|\tilde{x}_t) \end{aligned} \quad (61)$$

Deuker (2005) reports an application of his Qual-VAR to predicting the recession of 2001. Hence we look at the predictions of  $S_t$  in this context with the techniques just discussed, which take account of the fact that the states may not be known. The task is to forecast 2000/4-2003/3. Deuker assumes that  $S_t = 1$  in 2000/3. We recognize that this could not be known and so generate an estimate of  $S_t$  for 2000/3 using our procedure described above, although the results are much the same if we do in fact assume that  $S_t = 1$  for that period. It is necessary to have a way of generating  $g(x_t)$  over the

Table 1: Probability of a United States Recession, 200/4-2003/3

	AR(5)	VAR(5)
2000/4	.24	.24
2001/1	.36	.36
2001/2	.42	.45
2001/3	.42	.49
2001/4	.40	.49
2002/1	.37	.47
2002/2	.34	.44
2002/3	.30	.40
2002/4	.27	.36
2003/1	.24	.32
2003/2	.22	.29
2003/3	.20	.26

forecast period. To ensure comparability with Dueker we will make  $g(\cdot)$  a linear function.  $x_t$  will be the spread lagged two periods. One possibility is to produce forecasts of  $x_t$  using an AR(5) in  $x_t$ . Another is to use the broader set of variables in Deuker and take the forecasts of the spread to be from a VAR(5) in the GDP growth, inflation, the spread and an interest rate. We use both below.

Table 1 shows the probability of a recession over the period 2000/4-2003/3 given that the information used is just the knowledge of  $S_t$  in 2000/1 and 2000/2 along with the yield spread up to and including 2000/3 (as well as the other variables in the case of the VAR).

Comparing these to the results in Deuker (Table 2, p 100) we see that the pattern is the same but the latter reports a maximum probability of .57, which is a very high probability given such a weak recession (indeed Dueker notes that the Qual-VAR forecast of GDP growth never becomes negative). We were not able to replicate his Table 2 probabilities with the program Dueker supplied to us, getting a maximum probability of .55 and a probability in 2003/3 of .32, but these seem reasonably consistent. To understand why our estimates are smaller consider first the comparison between the AR and VAR. The VAR produces higher probabilities because it features three forecasts of the spread that are negative, with values of -.38, -.34 and -.02, whereas the AR never has any negative forecast, although there is one quar-

ter of a small positive value. Since Dueker utilizes a VAR in his work one would therefore expect a probability of at least .5. It should be said that the AR produces a much better forecast of the actual path of the spread than the VAR, as there are no negative spreads in the forecast period.

The other source of difference is that in 2003/3 the probability of a recession from the Qual-VAR is .29 which is high compared to the unconditional probability of .17 in the data. This suggests that the model has a tendency to assign a high probability to a recession. If we extend the forecast period for the AR and VAR models to 25 periods, the probability of a recession would be given as .174 i.e. both models return to the unconditional mean of  $S_t$  as the forecast. In contrast, simulating out the Qual-VAR produces an unconditional forecast of the probability of a recession of .39. One problem may be that in the simulations used to do the Bayesian forecasts unstable VARs were retained provided the maximum root was less than 1.02.. In a sense this is a specification test since the forecast should return to the unconditional mean in a stationary context, and  $S_t$  will be a stationary random variable. Thus the Qual-VAR does not seem to have this property and directs attention to the possibility of specification errors in it based on treating  $S_t$  is exogenous that were mentioned earlier.

## 5 Conclusion

We have made the argument that constructed states  $S_t$  require careful treatment if they re to be used in econometric work since they are very different in their nature to the binary states often modelled in micro-econometrics. One has to allow for the fact that they are essentially Markov Chains when engaging in a broad range of estimation and inference methods. But, to date, the nature of the  $S_t$  has mostly been ignored, with the potential for quite misleading estimates and inferences. We have suggested some methods to deal with this fact.

## 6 Appendix

The determination of these transition probabilities becomes much more complex with the “two quarters rule” as the conditioning event  $S_{t-1} = 1$  will place some restrictions upon the past sample paths for  $\{\Delta y_t\}$  that are associated

with an *ETS*. For example the sequence

$$\{\Delta y_{t+1}, \Delta y_t, \Delta y_{t-1}, \Delta y_{t-2}, \dots\} = \{-, -, -, +, \dots\} \quad (62)$$

would be incompatible with  $S_{t-1} = 1$  since the negative growth at  $t-1$  would match with the negative growth at  $t$  and so the expansion would have been terminated at  $t-1$ . It is clear that the sample paths  $\{\Delta y_{t-1}, \Delta y_{t-2}, \dots\}$  that are compatible with  $S_{t-1} = 1$  and  $\{\Delta y_{t+1} < 0, \Delta y_t < 0\}$  must have the form  $\{+, \dots\}$  and in such paths we must encounter a  $\{+, +\}$  before we encounter a  $\{-, -\}$ . If this did not happen e.g. we had for  $\{\Delta y_{t-1}, \Delta y_{t-2}, \dots\}$  the path  $\{+, -, +, -, -, \dots\}$ , then the recession would have begun at  $t-5$  and would still be running when we reach  $t-5$

Now let us consider an enumeration of the paths that are consistent with  $S_{t-1} = 1$ . This is done in the matrix below where the first column represents time and subsequent columns represent paths along which we are assured that  $S_{t-1} = 1$ . The notation used is as follows:

- “+” indicates  $\Delta y_t > 0$ ;
- “-” indicates  $\Delta y_t < 0$ ;
- “\*” before a “-” indicates that any pattern for the observations can occur along the path up to and including that point;
- “\*” following a “+” indicates that any pattern for the observations can occur along the path from that point forward.

Thus looking at the second column the “+, +” at  $t$  and  $t-1$  assures us that  $S_{t-1} = 1$  along all paths that exhibit this pattern at  $t$  and  $t-1$ . Similarly, the “-” at  $t$  and the “+, +” at  $t-1$  and  $t-2$  assures us that all paths with this pattern are consistent with  $S_{t-1} = 1$ . Similar logic can be applied to all the subsequent paths.

$$\begin{bmatrix}
t+1 & * & * & * & * & * & * & \cdots \\
t & + & - & + & - & + & - & \cdots \\
t-1 & + & + & - & + & - & + & \cdots \\
t-2 & * & + & + & - & + & - & \cdots \\
t-3 & & * & + & + & - & + & \cdots \\
t-4 & & & * & + & + & - & \cdots \\
t-5 & & & & * & + & + & \cdots \\
t-6 & & & & & * & + & \cdots \\
\vdots & & & & & & * & \ddots
\end{bmatrix} \tag{63}$$

To understand the derivation of these paths suppose we start with the four possible outcomes for  $(\Delta y_t, \Delta y_{t-1})$ , namely  $\{+, +\}$ ,  $\{-, +\}$ ,  $\{+, -\}$  and  $\{-, -\}$ . The last would give  $S_{t-1} = 0$  and the first  $S_{t-1} = 1$ ; thus the first becomes the second column of the table. The other two outcomes do not enable us to decide what the state for  $S_{t-1}$  is and so we proceed to observation  $t - 2$  and consider what happens to each of them as we add on a  $-$  or a  $+$ . Thus  $\{-, +, +\}$  will give  $S_{t-1} = 1$  and that becomes the third column. But  $\{-, +, -\}$  produces no resolution and one needs to proceed to  $t - 3$ . Augmenting  $\{+, -\}$  with a  $+$  also fails to resolve the indeterminacy while adding on a  $-$  result in  $S_{t-1} = 0$ . Consequently that path has to be continued on to  $t - 3$  as well. The process continues in this way and all columns of the matrix will eventually be enumerated by such a strategy.

To formalize the discussion it is helpful to separate the set of paths that are consistent with  $S_{t-1} = 1$  into two subsets. Let  $E_t$  be the set of paths such that  $\{\Delta y_t > 0 \text{ and } S_{t-1} = 1\}$  and  $F_t$  be the set of paths such that  $\{\Delta y_t < 0 \text{ and } S_{t-1} = 1\}$ . If we introduce the notation that

- $[+-]_t^j$  represents the fragment of the path along which there are  $j$  repetitions of the pattern in the  $[+-]$ .with the leading term in the pattern being located at time  $t$ ,
- $[++]_t$  represents the fragment of path where the pattern ” ++” occurs with the first ” +” being at  $t$  and the second at  $t - 1$
- $[-]_t$  represents the case where  $\Delta y_t < 0$ ,

the sets  $E_t$  and  $F_t$  can be enumerated as

$$E_t = \left\{ [++]_t; [+ -]_t [++]_{t-2}; [+ -]_t^2 [++]_{t-4}; \dots; [+ -]_t^j [++]_{t-2j}; \dots \right\} \quad (64)$$

$$F_t = \left\{ \begin{array}{l} [-]_t [++]_{t-1}; [-]_t [+ -]_{t-1} [++]_{t-3}; \\ [-]_t [+ -]_{t-1}^2 [++]_{t-5}; \dots; [-]_t [+ -]_{t-1}^j [++]_{t-2j-1}; \dots \end{array} \right\}. \quad (65)$$

Thus, using the notation that  $\Pr(E_t)$  represents the probability that the path is drawn from the set  $E_t$ , and recognizing that the paths are mutually exclusive, we have (to simplify notation we have omitted the conditioning on  $\mathfrak{S}_{t+1}$  in equations (66), (67), (68) and (70)).<sup>1</sup>

$$\Pr(E_t) = \sum_{j=0}^{\infty} \Pr\left([+ -]_t^j [++]_{t-2j}\right) \quad (66)$$

and

$$\Pr(F_t) = \sum_{j=0}^{\infty} \Pr\left([-]_t [+ -]_{t-1}^j [++]_{t-2j-1}\right). \quad (67)$$

By definition

$$\Pr(S_{t-1} = 1) = \Pr(E_t) + \Pr(F_t). \quad (68)$$

Interest also centres on the joint event  $\Pr\{S_t = 0, S_{t-1} = 1\}$ ; this will involve the set  $G_{t+1}$  defined as

$$G_{t+1} = \left\{ \begin{array}{l} [ - - ]_{t+1} [++]_{t-1}; [ - - ]_{t+1} [+ -]_{t-1} [++]_{t-3}; [ - - ]_{t+1} [+ -]_{t-1}^2 [++]_{t-5}; \dots \\ \dots; [ - - ]_{t+1} [+ -]_{t-1}^j [++]_{t-2j-1}; \dots \end{array} \right\} \quad (69)$$

Then, since  $S_t$  is a stationary process,

$$p_{10} = \frac{\Pr(S_t = 0, S_{t-1} = 1)}{\Pr(S_{t-1} = 1)} = \frac{\Pr(G_{t+1})}{\Pr(E_t) + \Pr(F_t)}. \quad (70)$$

---

<sup>1</sup>To simplify notation we have omitted the conditioning on  $\mathfrak{S}_{t+1}$  in equations (66), (67), (68) and (70).

If  $\Pr(S_t = 1, S_{t-1} = 0)$  is constant, which essentially requires  $\Delta y_t$  to be a random walk with time invariant drift and variance, then  $\Pr(S_t = 1, S_{t-1} = 0) = \Pr(S_t = 0, S_{t-1} = 1)$  (as the number of peaks and troughs must be the same). Using this in conjunction with  $\Pr(S_t = 0) = 1 - \Pr(S_t = 1)$  we can directly derive  $p_{01}$  from the same information as used to construct  $p_{10}$ . If  $\Pr(S_t = 1, S_{t-1} = 0)$  is time varying (as would be the case where  $\mu_t$  depends on some exogenous variable) then one also needs to enumerate the various paths where  $S_{t-1} = 0$ .

Considering the limits of  $E_t$  etc we get

$$\begin{aligned}\Pr(E) &= \sum_{j=0}^{\infty} (1-\psi)^2 [\psi(1-\psi)]^j \\ &= \frac{(1-\psi)^2}{1-\psi(1-\psi)}\end{aligned}\tag{71}$$

$$\begin{aligned}\Pr(F) &= \sum_{j=0}^{\infty} \psi(1-\psi)^2 [\psi(1-\psi)]^j \\ &= \frac{\psi(1-\psi)^2}{1-\psi(1-\psi)}\end{aligned}\tag{72}$$

$$\begin{aligned}\Pr(G) &= \sum_{j=0}^{\infty} \psi^2(1-\psi)^2 [\psi(1-\psi)]^j \\ &= \frac{\psi^2(1-\psi)^2}{1-\psi(1-\psi)}\end{aligned}\tag{73}$$

and so

$$p_{10} = \frac{\psi^2}{(1+\psi)}\tag{74}$$

$$p_{11} = \frac{1+\psi-\psi^2}{(1+\psi)}\tag{75}$$

$$p_{01} = \frac{(1-\psi)^2}{2-\psi}\tag{76}$$

$$p_{00} = \frac{1+\psi-\psi^2}{2-\psi}\tag{77}$$

Now in some of the literature we deal with it is assumed that the process for  $\Delta y_t$  depends linearly upon some other variable  $x_t$  in the following way:

$$\Delta y_t = a + bx_t + u_t \quad (78)$$

where the  $x_t$  are taken to be strictly exogenous (and so can be conditioned upon) and  $u_t$  is *n.i.d.*(0, 1). If  $\psi_t = \Phi(-a - bx_t)$ , applying the enumeration method results in

$$\begin{aligned} \Pr(E_{t+1}|\mathfrak{S}_{t+1}) &= \sum_{j=0}^{\infty} \Pr\left([+-]_{t+1}^j [++]_{t+1-2j}\right) \\ &= (1 - \psi_{t+1})(1 - \psi_t) \\ &\quad + \sum_{j=1}^{\infty} \left\{ \left[ \prod_{i=0}^{j-1} (1 - \psi_{t+1-i}) \psi_{t-i} \right] (1 - \psi_{t-2j+1})(1 - \psi_{t-2j}) \right\} \end{aligned} \quad (79)$$

and

$$\begin{aligned} \Pr(F_{t+1}|\mathfrak{S}_{t+1}) &= \sum_{j=0}^{\infty} \Pr\left([-]_{t+1} [+-]_t^j [++]_{t-2j}\right) \\ &= \psi_{t+1}(1 - \psi_t)(1 - \psi_{t-1}) + \\ &\quad \psi_{t+1} \sum_{j=1}^{\infty} \left\{ \left[ \prod_{i=0}^{j-1} (1 - \psi_{t-i}) \psi_{t-i-1} \right] (1 - \psi_{t-2j})(1 - \psi_{t-2j-1}) \right\}. \end{aligned} \quad (80)$$

Letting  $P_t = \Pr(S_t = 1|\mathfrak{S}_{t+1})$  under the two quarters rule gives

$$P_t = \Pr(E_{t+1}|\mathfrak{S}_{t+1}) + \Pr(F_{t+1}|\mathfrak{S}_{t+1}) \quad (81)$$

It is clear from this expression that the use of the two quarters dating rule means that  $P_t$  is a function not only of  $x_t$  but also of  $x_{t+1}$  and the entire past history of  $x_t$ . Moreover it does not have a single index form i.e. does not depend upon  $\alpha + x_t\beta$  alone. Only if the dating rule had been the ‘‘calculus’’ one would  $\Pr(S_t = 1|\mathfrak{S}_{t+1}) = (1 - \psi_t)$  be a function of  $x_t$  only. Clearly the lesson of this analysis is that one cannot just assume that  $\Pr(S_t = 1)$  is a function of a contemporaneous variable only; it is necessary that one know how the  $S_t$  were generated in order to be able to write down the correct likelihood.

## 7 References

Abiad, A. (2003), "Early Warning Systems: A Survey and a Regime-Switching Approach", *IMF Working Paper #32*.

Artis M.J. and W. Zhang W (1997), "International Business Cycles and the ERM". *International Journal of Finance and Economics*, 2,1, January.

Artis M.J. and W. Zhang, (1999), "Further Evidence on International Business Cycle and the ERM: is there a European Business Cycle?", *Oxford Economic Papers*, 51, 120-32.

Artis, M.J., Z.G., Kontolemis and D.R. Osborn (1997), "Business Cycles for G7 and European Countries". *Journal of Business*, 70, 249-79.

Birchenall, C.R., H. Jessen, D.R. Osborne and P. Simpson (1999), "Predicting U.S. Business-Cycle Regimes", *Journal of Business and Economic Statistics*, 17, 313-23.

Bordo, M, B. Eichengreen, D. Klingbiel and M. S. Martinez-Peria (2001), "Financial Crises", *Economic Policy*, 54-82.

Bordo, M.D. and D.C. Wheelock (2006), "When Do Stock Market Booms Occur? The Macroeconomic and Policy Environments of 20th Century Booms", *Federal Reserve Bank of St Louis Working Paper 2006-051A*

Brailsford, T., Heaney, R., A. and Shi, J., 2001, "The Cyclical Behaviour of the IPO market in Australia", *Accounting Research Journal*, 14, 1,17-34.

Bry, G., Boschan, C., (1971), *Cyclical Analysis of Time Series: Selected Procedures and Computer Programs*, New York, NBER.

Camba-Mendez, G. and D. Rodriguez-Palenzuela (2003), "Assessment Criteria for Output Gap Estimates", *Economic Modelling*, 20, 529-62.

Cashin, P.,C.J. McDermott and A. Scott (2002), "Booms and Slumps in World Commodity Prices", *Journal of Development Economics*, 69, 277-96.

Cashin, P. and C.J. McDermott (2002), "Riding on the Sheep's Back: Examining Australia's Dependence on Wool Exports", *Economic Record*, 78, 249-263.

Chin, D., Geweke, J., and P. Miller, (2000), "Predicting Turning Points", *Federal Reserve Bank of Minneapolis Research Department Staff Report No. 267*.

Diebold, F.X. and G. D. Rudebusch (1990), "A Non-Parametric Investigation of Duration Dependence in the American Business Cycle", *Journal of Political Economy*, 98, 596-616.

Dueker, M. (2005), "Dynamic Forecasts of Qualitative Variables: A Qual VAR Model of U.S. Recessions", *Journal of Business and Economic Statis-*

*tics*, 23, 96-104.

Durland, J.M. and T.H. McCurdy, T.H., (1994), "Duration-Dependent Transitions in a Markov Model of US GNP Growth". *Journal of Business and Economic Statistics*, 12, 279-88.

Eichengreen, B., A.K. Rose and C. Wyplosz (1995), "Exchange Rate Mayhem: The Antecedents and Aftermath of Speculative Attacks", *Economic Policy*, 21, 251-312.

Estrella, A. and F.S. Mishkin (1998), "Predicting US Recessions: Financial Variables as Leading Indicators", *Review of Economics and Statistics*, LXXX, 28-61.

Hamilton, J.D., (1989), "A New Approach to the Economic Analysis of Non-Stationary Times Series and the Business Cycle", *Econometrica*, 57, 357-84.

Harding D., and A.R. Pagan, (2002), "Dissecting the Cycle: A Methodological Investigation", *Journal of Monetary Economics*, 49(2), 365-81.

Harding, D. and A. R. Pagan (2003) "A Comparison of Two Business Cycle Dating Methods" *Journal of Economic Dynamics and Control*, 27, 1681-90.

Harding D., and A.R. Pagan, (2006), "Synchronisation of Cycles", *Journal of Econometrics* (In Press)

Ibbotson, R.G., and J.J. Jaffee (1975), " 'Hot Issue' Markets", *Journal of Finance*, 30, 1027-1042.

Kaminsky, G.I. and C.M. Reinhart (1999), "The Twin Crises: The Causes of Banking and Balance-of- Payments Problems", *American Economic Review*, 89, 473-500.

Kedem, B., (1980), *Binary Time Series*, Marcel Dekker, New York.

Lunde, A. and A.Timmermann (2004), "Duration Dependence in Stock Prices: An Analysis of Bull and Bear Markets", *Journal of Economic and Business Statistics*, 22, 253-73.

Neftci, S.N. (1984), "Are Economic Times Series Asymmetric over the Business Cycle", *Journal of Political Economy*, 92, 307-28.

Newey, W.K. (1984), "A Method of Moments Interpretation of Sequential Estimators", *Economics Letters*, 14, 201-06.

Newey, W.K. and K. West (1994), "Automatic Lag Selection in Covariance Matrix Estimation", *The Review of Economic Studies*, Vol 61, Issue 4, October.

Ohn, J., L. Taylor and A.R. Pagan (2004), "Testing for Duration Dependence in Economic Cycles", *The Econometrics Journal*, 7, 528-49.

Pagan, A.R. and K. Sossounov (2003), "A Simple Framework for Analysing Bull and Bear Markets", *Journal of Applied Econometrics*, 18, 23-46.

Pesaran, M.H. and A. Timmermann (1992), "A Simple Nonparametric Test of Predictive Performance", *Journal of Business and Economic Statistics*, 10(4), 461-65.

Raftery, A.E (1985), "A Model for High-Order Markov Chains", *Journal of the Royal Statistical Society, Series B*, 47, 528-539.

Robinson, P. (1988), "Root-N-Consistent Semiparametric Regression", *Econometrica*, 56, 931-54.

Sichel, D. E. (1994). "Inventories and the Three Phases of the Business Cycle", *Journal of Business and Economic Statistics*, 12, 269-77.