ELSEVIER Journal of Monetary Economics 41 (1998) 455–473

JOURNAL OF Monetary ECONOMICS

Money and output viewed through a rolling window

Norman R. Swanson*

Department of Economics, Pennsylvania State University, University Park, PA 16802, USA

Received 21 April 1995; received in revised form 16 September 1997; accepted 2 October 1997

Abstract

We examine the extent to which fluctuations in the money stock anticipate (or Granger cause) fluctuations in real output using a variety of rolling window and increasing window estimation techniques. Various models are considered using simple sum as well as Divisia measures of $M1$ and $M2$, income, prices, and both the T-bill rate and the commercial paper rate. Findings indicate that the relation between income, money, prices, and interest rates is stable, as long as sufficient data are used, and that there is cointegration among the variables considered, although cointegration spaces become very difficult to estimate precisely when smaller windows of data are used. Further, both $M1$ and $M2$ are shown to be important predictors of income for the entire period from 1960:2–1996:3, based on *modified* versions of what we term the 'most damaging' specifications from Friedman and Kuttner (1993) and Thoma (1994). Our new evidence is based in part on a rather novel model selection approach to examining the relationship between money and income.

JEL classification: C22; C32; C53; E52

Keywords: Schwarz information criterion; Cointegration; Prediction; Causation

1. Introduction

An issue of continuing interest among economic theorists and policy makers is the extent to which fluctuations in the money stock anticipate fluctuations in real income. Since the reconsideration of monetarism by Sims (1980), a great

* E-mail: nswanson@psu.edu

0304-3932/98/$19.00
PII S0304-3932(98)00005-1

number of papers have attempted to answer the question: 'Does money matter?'. For instance, Christiano and Ljungqvist (1988) find significant Granger-causality using a bivariate money-income model, while Stock and Watson (1989) include interest rates and prices, and also find evidence of the forecasting ability of money for income, when money is measured by quadratically detrended M1. Recently, however, Friedman and Kuttner (1993) used Stock and Watson's specification to show that extending the sample through 1990 causes money to become insignificant in the income equation. Further, by replacing the Treasury bill rate with the commercial paper rate, or by including an interest rate spread, money is also seen to be insignificant, even when Stock and Watson's sample period (1960:2–1985:12) is used. Thoma (1994) also finds little evidence of money-income causality, in the context of rolling regression estimates based on increasing samples of data. Although these results are certainly not exhaustive of the recent record on money-income causality (see e.g. McCallum, 1979; Bernanke, 1986; Blanchard and Quah, 1989; King et al., 1991; King and Watson, 1992; Stock and Watson, 1993; Hafer and Kutan, 1997), it is noteworthy that they are all based on vector autoregressions which use nominal simple sum money measures, do not include long-run cointegrating restrictions, and are estimated using either 6 or 12 lags of each endogeneous variable.

In this paper we re-examine the predictive content of money for income by using Divisia as well as simple sum $M1$ and $M2$ aggregates to measure money. We also ask the question, 'Are money, income, prices, and interest rates linked in the long-run by the presence of common stochastic trends? If so, then how many common stochastic trends are present in the data, and how does this impact on the marginal predictive content of money for income?' We also examine the impact on earlier findings of specifying the number of lags in regression models by using model selection criteria, rather than fixing the number of regressors a priori. Further, we attempt to differentiate between tests which are designed to examine *in-sample* coefficient restrictions, and those which are designed to select optimal *out-of-sample* forecasting models. This is of some interest, as models which have statistically significant *in-sample* monetary components are not guaranteed to provide optimal *forecasts* of income, relative to simpler models which do not contain monetary components, and vice versa, for example.

The approach which we take is to use rolling 10 and 15 year fixed and increasing windows (samples) of data to estimate the relationship between money and income. This is done for a number of reasons. First, by using fixed windows, we allow that the system may be evolving over time. Thoma (1994) uses a growing window of data, as he is not concerned with tracking a possibly evolving system in the sense of time-varying parameters, but rather assumes that the system is evolving to some final form. Second, our approach addresses the issue of sub-sample instability by considering: (i) a sequence of 315 different

samples (starting with 1960:2–1970:1 and ending with 1986:4–1996:3) for the 10 year fixed window, (ii) a sequence of 315 different samples (starting with 1960:2–1970:1 and ending with 1960:2–1996:3) for the 10 year increasing window, and (iii) a sequence of 255 samples for the 15 year fixed window. Third, our approach allows us to examine the stability of cointegrating vector and cointegrating space rank estimates throughout a sequence of samples in a possibly evolving system. The paper is perhaps closest to that of Hafer and Kutan (1997). However, our approach differs from theirs in a number of respects. We use monthly data, and find that both $M1$ and $M2$ are important for predicting income. Hafer and Kutan, on the other hand, use quarterly and annual data, and find little evidence that $M1$ Granger-causes income. Some other features that differentiate our analysis from Hafer and Kutan's are that we: consider fixed as well as increasing rolling windows of data; include trending characteristics which agree with the findings of Stock and Watson (1989); and use a causality test which is validly applied to data which are stationary, integrated, cointegrated, or any combination thereof.

Our empirical findings suggest that models which include money, income, prices, and interest rates do contain common stochastic trends (see Stock and Watson, 1993; Hafer and Kutan, 1997 for related evidence). Further, we note that the rank of the cointegrating space is unstable when 10 and 15 year fixed windows of data are used. This evidence is consistent with the finding of Stock and Watson, 1993 that elasticities based on money demand models which use money, income, prices, and interest rates are quite sensitive to the final regression date, when monthly post-war data are used. Interestingly, though, the estimated cointegrating relations stabilize dramatically when increasing windows of data are used, as the sample size is increased. Also, the Treasury bill – commercial paper spread used as a regressor by Friedman and Kuttner (1993) arises naturally as a cointegrating vector in a number of the systems which we examine. Overall, by accounting for cointegration, selecting the number of lags based on Schwarz and Akaike information criteria (SIC and AIC, respectively), and examining Granger causality tests which directly address the issue of predictive ability, we find surprisingly robust new evidence of the marginal predictive content of money for income.

By adopting a rolling window approach, we believe that we contribute not only to the discussion of the usefulness of money as a predictor of future income, but also to the methodology of examining this and similar issues. One dimension of this contribution is that we consider models which incorporate cointegration of unknown form, thus avoiding the problem of estimating potentially unstable cointegrating relations. We also differentiate between *forecasting* models, and models which are designed for in-sample inference (e.g. to test economic theories). Contributions are also attempted in several other related interesting areas. For example, we examine the relative performance of the SIC and the AIC for selecting lags, in the context of constructing ex ante one-step

ahead forecasting models. We do not, however, offer evidence which contrasts credit and money views of the monetary transmission mechanism. This is because both credit and money views allow for movements in market interest rates, *in the same direction*, and we use market interest rates in our models. In order to differentiate between the two theories, one could, however, use the interest rate on loans (although such data is currently unavailable), or some credit aggregate. The rest of the paper is organized as follows. Section 2 discusses the data, while Section 3 outlines the models considered. Estimation and testing procedures are summarized in Section 4, while Section 5 presents our empirical result. Section 6 contains a summary and concluding remarks.

2. The data

The variables used are the same as those examined by Christiano and Ljungqvist (1988), Stock and Watson (1989), Friedman and Kuttner (1993), Thoma (1994), and others. In particular, monthly observations on the log of seasonally adjusted nominal $M1$ ($m1_t$), the log of seasonally adjusted nominal $M2$ ($m2_t$), the log of industrial production (y_t), the log of the wholesale price index (p_t), the secondary market rate on 90-day U.S. Treasury bills (i_t), and the interest rate on six-month dealer-placed prime commercial paper (c_t) are used. The sample period considered is January 1959–March 1996.

A feature which differentiates the simple sum measures of money which we examine is that outside money – the monetary base – accounts for less than 10% of the broader $M2$ aggregate. Also, our $m2_t$ series exhibits erratic behavior since 1985, which can be accounted for by well documented recent shifts in the public's demand for money balances. This probably accounts at least in part for recent evidence that the relationship between $m2_t$,y_t, and p_t has been unstable in recent years. Our approach for dealing with shifting money demand is to consider Divisia monetary aggregates in addition to $m1_t$ and $m2_t$. In particular, we use logged Divisia monetary aggregates for $M1$ ($dm1_t$), $M2$ ($dm2_t$), and $M3$ ($dm3_t$). The Divisia monetary aggregates were obtained from the St. Louis Federal Reserve Bank, are described in Anderson et al. (1996), and are based on the work of William Barnett (see e.g. Barnett (1978), Barnett (1980), Barnett (1990), and the references contained therein).

3. Empirical methodology

3.1. Previous preferred models

Our first objective is to characterize what we shall term the 'preferred' specifications used in a number of previous money-income analyses. By 'preferred' we

mean the model which we interpret as best illustrating the authors' main findings. In order to do this, we start by assuming that there is no cointegration among money, income, prices and interest rates. This allows us to consider models close to those estimated by Friedman and Kuttner (1993), Stock and Watson (1989), and Thoma (1994). However, our models still depart from the specifications used by the above authors in at least two ways. First, we linearly as well as quadratically detrend our data. This strategy is adopted because some authors assume that money is linearly detrended, while others assume that it is quadratically detrended. Second, for the remainder of the paper, the number of lags used in our models is estimated by selecting specifications which minimize Schwarz and Akaike Information Criteria. This is contrary to the frequent approach of arbitrarily fixing the number of lags at 6 or 12. Along these lines, vector autoregression (VAR) models of the following form are specified:

$$\Delta x_t = \gamma_0 + \tau(t) + C(L)\Delta x_{t-1} + \varepsilon_t, \tag{1}$$

where ε_t is a vector of innovations, $\tau(t)$ is a polynomial function of time (where $\tau(t) = 0$ or $\tau(t) = \gamma_1 t$), and $C(L)$ is a matrix polynomial in the lag operator L.[1] The vector x_t is either $(m_t, y_t, p_t, i_t)'$ or $(m_t, y_t, p_t, i_t, c_t)'$, where m_t is alternately $m1_t$, $m2_t$, $dm1_t$, $dm2_t$, or $dm3_t$.[2] The model with $x_t = (m1_t, y_t, p_t, i_t)'$ and $\tau(t) = \gamma_1 t$ is assumed to be the preferred model of Stock and Watson (1989), while Thoma (1994) prefers the same model, but with $\tau(t) = 0$. Friedman and Kuttner (1993) prefer the model with $x_t = (m1_t, y_t, p_t, i_t, c_t)'$ and $\tau(t) = \gamma_1 t$.[3] (It should be noted, though, that Stock and Watson (1989) and Friedman and Kuttner (1993) consider various different polynomial functions of time in their models.) In all, the 40 models ($2x_t$ vectors $\times$ 5 money measures $\times$ 2 trends specifications $\times$ 2 lag order selection criteria) which are estimated based on Eq. (1) include not

[1] In passing, it is worth mentioning that the linear and fixed parameter vector autoregression methodology which we adopt is subject to a variety of reservations. For example, time varying parameter and other sorts of nonlinear models are receiving increasing attention in the literature (see e.g. Granger and Terasvirta (1993); Potter, 1995; and the references contained therein).

[2] Based on Augmented Dickey–Fuller (ADF) tests for one and two unit roots, and based on regressions of the first difference of each series against a constant, time and six of its own lags, we concluded that our data are consistent with the following specifications: Δy_t and Δp_t are $I(0)$ with drift; $\Delta m1_t$, $\Delta m2_t$, $\Delta dm1_t$, $\Delta dm2_t$, and $\Delta dm3_t$, are all $I(0)$ with small but statistically significant determinsitic trends; and c_t and i_t are $I(1)$ with no drift. Even though the nonnegativity of our interest rate series raises some conceptual difficulties with our characterizations, these findings are consistent with the specifications used by Friedman and Kuttner (1993), Stock and Watson (1989), and Thoma (1994), among others.

[3] More precisely, the 'most damaging' model of Friedman and Kuttner (1993) is assumed to be Eq. (1) with $x_t = (m1_t, y_t, p_t, i_t, c_t)'$, $\tau(t) = \gamma_1 t$, and with additional regressors which are error-correction terms that incorporate cointegrating restrictions in the VAR model. It turns out that one of the error-correction terms which we specify is the T-bill – commercial paper spread, in agreement with Friedman and Kuttner's use of the same variable (see below discussion).

only the 'preferred' specifications from previous studies, but also a variety of other models.[4]

3.2. Stochastic trending properties of the data

Standard F-tests or Wald-tests for Granger causality are prone to severe upward size distortions when vector error correction (VEC) models are estimated using only differenced data, without accounting for cointegrating restrictions. One of the reasons why this problem arises is that the moving average representation for a model with cointegrated regressors will not yield a finite order VAR representation. Put another way, testing bias arises in part because least squares becomes 'confused' when a potentially significant variable (the error correction term) is omitted from the regression model. Interestingly, Stock and Watson (1989) find no evidence of cointegration among y_t, $m1_t$, p_t and i_t, and Friedman and Kuttner (1993) and Thoma (1994), among others, adopt this finding when specifying their models. However, Friedman and Kuttner (1993) do include the c_t-i_t spread as one of their explanatory variables. This variable is characterized as $I(0)$, and thus may be interpreted as a cointegrating variable. Further, Stock and Watson (1993) find evidence of cointegration among our monthly series when they construct money demand equations in order to examine income and interest rate elasticities, and Hafer and Jansen (1991) find evidence of cointegration between real money balances, real income, and short-term interest rates using quarterly data from both 1915 and 1953 to 1988, when $M2$ is used to measure the aggregate stock of money. However, Hafer and Jansen find no cointegration when $M1$ is used in place of $M2$.

Given this rather mixed evidence concerning cointegration, we begin our analysis of the stochastic trending properties of our dataset by re-examining the data and sample period (1960:2–1985:12) considered by Stock and Watson (1989). The results of cointegration tests based on various trend specifications, and 6 and 12 lag VEC models with the variables $m1_t$, y_t, p_t, and i_t are available upon request, and can be summarized as follows. At a 1% significance level, trace test statistics support the presence of one cointegrating (CI) vector when the data are linearly detrended, and when an intercept, or an intercept and a trend is included in the CI relation, for models estimated with both 6 and 12 lags. However, when a quadratic trend is included in the levels data (as in the Stock and Watson 'linearly detrended nominal money growth' case) the 6 lag models all accept a null hypothesis of no cointegration, while the corresponding

[4] Our approach of examining models which use various different money measures is contrary to the common practice of considering only $m1_t$, and this use of alternate measures of money raises a number of important questions ranging from: 'What is the impact of different monetary policies?', to 'What impact can the Federal Reserve Board exert over broader money aggregates?' However, discussion of these and related questions is left to future research.

12 lag models reject the same null hypothesis. Since Stock and Watson include a quadratic trend in the levels data, and consider only 6 lags, our results agree with theirs. However, it is nevertheless interesting that when the model is specified with 12 lags, there appears to be strong evidence in favor of one CI vector. A final wrinkle is added to the picture, as the null hypothesis of at most one cointegrating relation is rejected in favor of 2 cointegrating relations at the 5% significance level for the model in which the following three conditions hold: (i) there are 12 lags, (ii) the levels data are linearly detrended, and (iii) the CI vector is constructed with an intercept and a deterministic trend. Thus, although the evidence remains mixed, it suggests that there may indeed be cointegration among m_t, y_t, p_t, and i_t, and that quite often one cointegrating restriction adequately characterizes the data. This hypothesis is further supported by noting that when VEC models with quadratically detrended $m1_t$, y_t, p_t, and i_t are estimated under the assumption that there is one CI restriction, the error-correction term is always significant at a 5% level in the y_t (and i_t) equation, regardless of whether the lag order is 6 or 12. Of final note is that when $m2_t$, $dm1_t$, $dm2_t$, or $dm3_t$ are used in place of $m1_t$, similar evidence in favor of cointegration among the four variables is found.

Given our somewhat mixed results concerning the number of cointegrating restrictions among the variables, however, the precise form of the basis of the cointegrating space remains uncertain. In order to examine this issue further, we estimated the rank and basis of the cointegrating space for 10 and 15 year fixed, as well as 10 year increasing length windows of data using Johansen's (1988), Johansen's (1991) method. In this way, we obtained 315 different cointegrating rank estimates, for example, when 10 year fixed windows of data were considered. The experiments were done for models using m_t, y_t, p_t, and i_t as well as for models using m_t, y_t, p_t, i_t, and c_t, where m_t was alternately $m1_t$, $m2_t$, $dm1_t$, $dm2_t$, and $dm3_t$. Although the results are far too numerous to list here, they are similar across all money measures, and are summarized in Figs. 1 and 2, where estimation results based on the use of $m2_t$ as the money measure are plotted.[5]

Fig. 1, Panel A contains plots of estimated cointegrating space ranks and lag orders (based on the SIC) for models estimated with increasing windows of data. Note that the cointegrating rank seems to stabilize at two when the data are linearly detrended, and at one when the data are quadratically detrended, when $m2_t$, y_t, p_t, and i_t are modeled. Thus, cointegrating rank estimates are dependent on the assumed deterministic trend components in the model. This problem is not disimilar from the difficulty in testing for trend versus difference stationarity in univariate series, and suggests that care needs to be taken at the initial model

[5] Note that the right vertical axes of the graphs, which are used to denote the number of estimated lags for each rolling regression model, run from 0 to 15 in Fig. 1, and from −30 to 15 in Fig. 2. This is done in order to ensure that the plotted lines in each graph do not overlap.

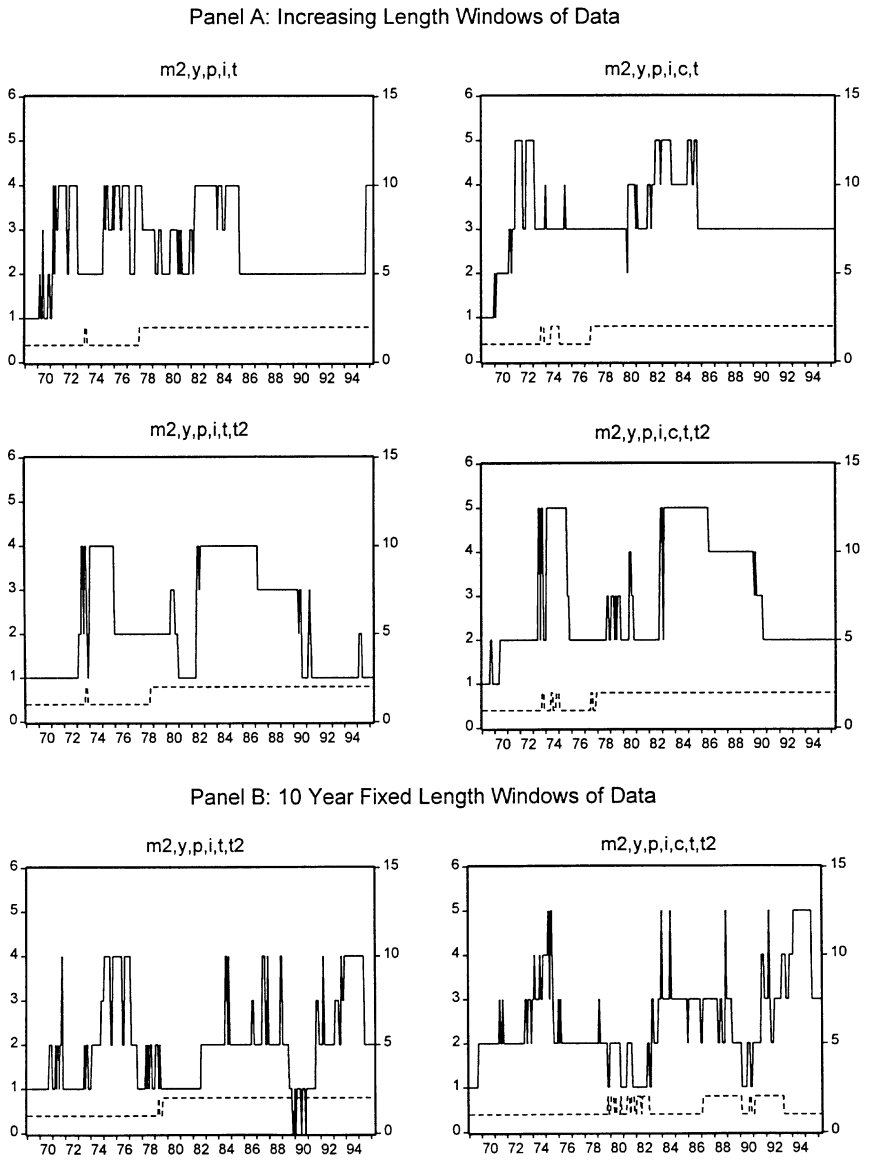

Fig. 1. Rolling estimates of lag orders (selected using the SIC) and cointegrating ranks.
Notes: All figures display estimates based on vector error correction models with increasing (Panel A) and fixed (Panel B) windows of data, starting with the sample period 1959:1–1968:12, and ending with the period 1959:1–1996:3 (for Panel A) and 1986:4–1996:3 (for Panel B). Dotted lines are estimated lag periods (and correspond to the right vertical axes), while solid lines are estimated cointegrating ranks (and correspond to the left vertical axes). Horizontal axes correspond to estimation end periods. t and $t2$ denote models estimated with linearly (t only) and quadratically (t and $t2$) detrended data. Variable names used in estimations are given above each graph.

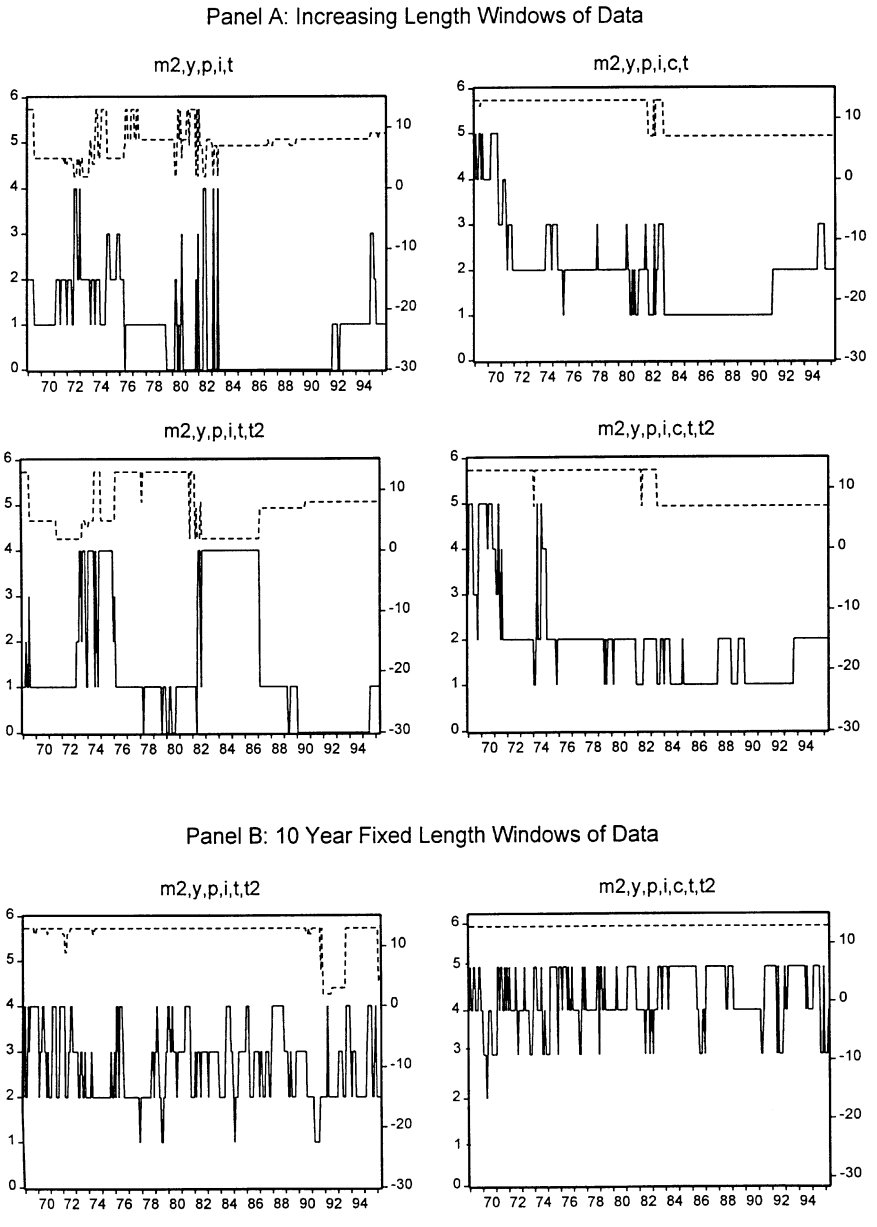

Fig. 2. Rolling estimates of lag orders (selected using the AIC) and cointegrating ranks.

specification stage of our analysis. Lag selection is not dependent on whether the data are linearly or quadratically detrended. However, our models appear to become much more confused when the AIC is used in place of the SIC to select lags (see Fig. 2, Panel A), with respect to both cointegrating rank estimation and lag selection. This suggests that stable models (with respect CI space rank and lag order) are more easily specified when the more parsimonious SIC is used. When the commercial paper rate is added to our models, the cointegrating rank increases by one, in all cases. This is interesting, as it suggests that the paper-bill spread used by Friedman and Kuttner (1993) is a candidate error-correction term. This conjecture is confirmed rather easily in the context of the five variable VEC model with lags selected by the SIC, with quadratically detrended data, and with $m2_t$ used as the money measure, for example, as each of the 75 most recent increasing windows of data (i.e. samples from 1960:2–1990:1 increasing monthly to 1960:2–1996:3), one of the two cointegrating vectors is c_t-i_t, based on a likelihood ratio test (see Johansen, 1988, Johansen, 1991). Of further note is that the null hypothesis that the other CI vector is $(1\ \ -1\ \ -1\ 0\ 0)$ almost always fails to reject, although confidence intervals are quite wide relative to those for the interest rate spread CI vector. (Actual parameter estimates are available upon request.)

Panel B of Figs. 1 and 2 contains similar plots for 10 yr fixed windows of data, but only for linearly detrended data (results for quadratically detrended data are similar). It is clear that the cointegrating rank evolves quite sporadically, and does not stabilize over time, regardless as to whether the SIC or the AIC is used to select the number of lags.[6] Given the above discussion, we conclude that the 120 month samples are too short to accurately estimate cointegrating spaces and ranks. Interestingly, this is the case even when the order of the VEC model is only one or two (i.e. when the SIC is used to choose the lag length), suggesting that in addition to a degrees of freedom (or confidence interval) problem, our shorter samples simply do not contain enough dynamic information to accurately estimate long-run relationships among the variables. When 15 year fixed windows of data are used, the picture that emerges is much the same as that illustrated in Panel B of Figs. 1 and 2. There are two possible remedies to this apparent difficulty in pinning down the cointegration in our fixed window models. First, longer samples of data can be used (e.g. use increasing windows of data). Second, a method can be used to test for the marginal predictive content of money for income which does not require estimation of the cointegrating restrictions in the system. The second approach is particularly appearing when it

[6] In one sense, these results are not surprising, given the evidence found by Gonzalo (1994) and Bewley et al. (1994), who discuss the small-sample tendency for maximum likelihood based procedures to generate biased and leptokurtic distributions. Also, it is worth noting that empirical analysis has shown that estimates of cointegrating spaces may differ, depending on which method of estimation is used (e.g. see Stock and Watson, 1993).

is suspected that the system may be evolving over time, so that even the use of increasing windows of data may not enable us to precisely estimate CI vectors and CI space ranks. In the next section we discuss these alternative testing approaches, and also consider a model selection based approach as an alternative to in-sample based Granger causality testing.

4. Estimation and testing

Least squares is used to estimate the parameters of the 40 VAR models given by Eq. (1), while maximum likelihood is used to estimate 40 related VEC models given by the following equation:

$$\Delta x_t = \beta_0 + \tau(t) + B(L)\Delta x_{t-1} + \sum_{i=1}^{r} \beta_i z_{i,t-1} + v_t, \tag{2}$$

where x_t and $\tau(t)$ are as discussed above, v_t is a vector of innovations, $z_{i,t-1} = \hat{\alpha}' x_{t-1}$, $i = 1, \ldots, r$, is a vector of $I(0)$ error-correction terms estimated in standard fashion using the methods proposed by Johansen (1988), Johansen (1991), r is the estimated rank of the cointegrating space, and $B(L)$ is a matrix polynomial in the lag operator L. As discussed above, the lag order of our models, say l, is chosen alternately using the SIC and the AIC. In all cases, the number of lags for each endogenous variable in the system is the same. At this point, it should perhaps be reiterated that Eqs. (1) and (2) represent models for a single fixed sample period. As we are implementing a rolling window approach, however, estimates of each of the parameters in Eqs. (1) and (2), including the coefficients, l, and r, may vary from sample to sample. For example, for 10 year fixed windows of data, 315 different sets of parameters are estimated. Corresponding to our above examination of rolling cointegration rank and lag length estimates, 10 and 15 year fixed as well as 10 year increasing windows of data are used in order to assess the extent to which fluctuations in money anticipate fluctuations in industrial production. However, given our findings with regard to the instability of cointegrating rank estimates, we wish to stress that the fixed window and the increasing window methods may yield quite different empirical findings, particularly given our lack of knowledge concerning how rapidly the economy is evolving over time.[7]

Three different types of causality tests are constructed, the first two of which are based on classical hypothesis testing principles. The first of these is the

[7] Granger (1996) points out that structural instability may be the most important problem facing forecasters today. We take the approach of allowing for possible structural instability by using rolling windows.

standard Wald test which is designed for stationary and difference stationary data. This test statistic is constructed for all models estimated using Eq. (1). The second test is the surplus lag regression test due to Dolado and Lütkepohl (1996) as well as Toda and Yamamoto (1995). This is also a standard Wald test, but is based on a VAR model in levels, and does not require pre-testing for cointegration, as models estimated in levels implicity include cointegration of unspecified form – a property which is not shared by difference VAR models. The test has the property that nonstandard asymptotics that arise when standard Wald tests are applied to integrated and/or cointegrated variables are forced in 'surplus' coefficient matrices. The only requirement for the consistency of this test is that the number of surplus lags 'added' to the original VAR model is equal to or greater than the largest order of integration of any of the variables in the system. Swanson et al. (1996) show that the surplus lag test has excellent finite sample size properties, particularly relative to standard Wald tests when the variables in the system are cointegrated. However, due to the inefficient estimation of systems with surplus lags, the finite sample power of the surplus lag test should be suspect. Further, a careful examination of the power functions of these types of tests still remains to be done, in particular to ensure that the local power is not flat in most or all directions. Nevertheless, recall that our estimates of cointegrating space ranks based on 10 and 15 year fixed windows of data are very erratic. This may be interpreted as evidence of the imprecise nature of these small sample estimates, and supports the use of surplus lag regression tests, which avoid the necessity of constructing cointegrating space estimates. Indeed, the loss of power associated with inefficient estimation of surplus lags may be less than the loss of power associated with the imprecise estimation of cointegrating spaces, were standard tests to be used. (Swanson et al., 1996 provide preliminary evidence that this tradeoff is important when constructing Granger causality tests.)

Both of our classical Wald tests are based on testing for the significance of money in the income equation. However, Granger causality tests have a natural interpretation as tests of predictive ability. In this sense, it may seem natural to use the results of causality tests when constructing forecasting models. However, any good forecasting model should be 'tested' using some sort of ex ante forecasting experiment before being implemented in any practical way. Along these lines, a reasonable alternative 'test' for non-causality could be to consider a set of different forecasting models, both with and without the variable whose causal effect is being examined. These 'competing' models could then be subjected to an ex ante forecasting analysis and the 'best' model chosen using some sort of out-of-sample model selection criterion. Then, if the 'best' model contains the variable of interest, noncausality is 'rejected'. Here, we consider a related alternative to classical Wald tests which is based on model selection, noting that Granger et al. (1995) suggest that although standard hypothesis testing has a role to play in terms of testing economic theories, it is more difficult to justify

using standard hypothesis tests for choosing between two competing models. One reason for their concern is that one model must be selected as the null, and this model is often the more parsimonious model (as in our case). However, it is often difficult to distinguish between the two models (because of multicollinearity, near-identification, etc.), so that the null hypothesis may be unfairly favored. For example, it is far from clear that pre-test significance levels of 5% and 1%, say, are optimal. The use of model selection criteria neatly avoids the problem of how to arbitrarily choose significance levels. Another advantage of the model selection approach is that it does not require specification of a correct model for its valid application. Further, the probability of selecting the truly best model approaches one as the sample size increases, if the model-selection approach is properly designed. This is contrary to the standard practice of fixing a test size, and rejecting the null hypothesis at that fixed size, regardless of sample size. Swanson and White (1995, 1996) discuss these and related features of model selection, while Swanson (1996) show using Monte Carlo that predictive accuracy tests based on the AIC and the SIC have empirical probabilities associated with selecting the wrong model which approach zero very quickly, even for sample sizes of 100, 200, and 300 observations. Our approach is to use two complexity based likelihood criteria, the AIC and the SIC, where

$$AIC = T\log|\hat{\Sigma}| + 2f \quad \text{and} \quad SIC = T\log|\hat{\Sigma}| + f\log(T),$$

where f is the total number of parameters in the system (if we are measuring the causal effect of some variable(s) on a *group* of more than one other variable), or, f is the number of parameters in the single equation of interest (when the *group* being examined is a single variable). Similarly, $\hat{\Sigma}$ is some standard estimate of the error covariance matrix, which is scalar if only one equation in the system is being examined. The AIC and SIC type tests (which we hereafter refer to as 'predictive accuracy' tests) are implemented as follows. The statistics are calculated for VEC versions of the models both with and without money variables.[8] If the 'best' model contains any money variable(s), then we have direct evidence of the marginal predictive content of money for income.[9]

[8] For each rolling window of data, we use a standard χ^2 hypothesis test on the estimated cointegrating vectors in the system to determine whether money has a nonzero weight in the cointegrating vector(s). This in turn allows us to determine which cointegrating vectors to include in our more parsimonious model which excludes money.

[9] Thus far, we have not differentiated between 'short'- and 'long-run' predictive ability. However, if we view cointegrating restrictions as corresponding to long-run relationships (see e.g. Granger and Lin, 1995), and lagged difference variables as corresponding to short-run relationships (as has been done elsewhere), then our difference variable Wald tests might be viewed as testing for short-run predictive ability, while the rest of our tests could be viewed as simultaneously testing for both short- and long-run predictive ability.

5. Empirical results

The results of the Granger causality and predictive accuracy tests are presented in Table 1 for the cases where money is measured using $m1_t$ and $m2t$. A similar table based on $dm1_t$, $dm2_t$, and $dm3_t$ has not been included because the results are numerically very similar to those presented in Table 1. (The table is available upon request.) Table 1 is broken into four groups of six rows, each corresponding to a different choice of variables. In turn, each set of variables is fit to VAR and VEC models using linearly and quadratically detrended data, 10 year fixed, 15 year fixed, and 10 year increasing windows of data, and based on lag orders chosen using both the AIC and the SIC. Entries in the table are rejection frequencies of the Granger noncausality null (for standard Walt tests and surplus lag Wald tests, which we call our 'classical' hypothesis tests), and the percentage of times that the 'best' model contains money (for the predictive accuracy tests).

Consider first the results based on standard Wald tests, which we call our 'difference Wald tests'. When the AIC is used to select lags, rejection frequencies range from 63% to 100%, regardless of window type and trend specification, for all models with $m2_t$. However, when $m1_t$ is used, rejection frequencies are much higher for the 10 year fixed window (around 65%) than for the other windows (where frequencies are often as low as 15%). This may suggest that models with $m1_t$ are evolving quite rapidly, so that the longer windows are less apt to effectively capture the money-income relationship. On the other hand, a degrees of freedom argument may also account for the apparent strength of the 10 year fixed window rejection frequencies, particularly as the AIC tends to overselect the lag order. The degrees of freedom argument is made even more believable by noting that rejection frequencies skyrocket to near unity for all longer windows when the SIC is used to select lag order. Paradoxically, though, the only low rejection frequencies for the cases where SIC is used are for $m2_t$ and the 10 year fixed window! In one sense, this constitutes evidence which is rather the opposite of when the AIC is used, at least with respect to $m2_t$. Thus, our evidence is rather mixed based on difference Wald tests. However, we know that the variables are cointegrated, and hence test bias may be driving our results, particularly for the cases where lag orders are chosen using the SIC, given that the SIC selects more parsimonious models than the AIC. This is one of the reasons why we also estimate surplus lag Wald tests. As mentioned above, these tests are designed to incorporate cointegration of unspecified form at the model estimation stage, and are based on levels VAR models. Because it is so important to include sufficient 'surplus' lags (in order that standard asymptotic critical values can be used), the AIC may be preferable for selecting lag order for these tests. Interestingly, cursory examination of the rejection frequencies in the AIC columns suggests that rejection rates are high only for 10 year fixed window cases. Since the 10 year fixed window cases are suspect (given the above discussion, and due to

Table 1
Summary of causality and predictive accuracy tests[a]
5% and 10% rejection rates and model selection results based on simple sum money measures

Model	Window	Deter-ministic trend	Av(SE) of estimated lags		Av(SE) of estimated CI rank		Difference Wald tests				Surplus lag Wald tests				Predictive accuracy test	
							5%	10%	5%	10%	5%	10%	5%	10%		
			AIC	SIC	AIC	SIC	AIC		SIC		AIC		SIC		AIC	SIC
m1,y,p,i	INCR	*t*	9.6(3.7)	1.6(0.5)	1.1(0.8)	1.9(0.6)	0.13	0.23	0.88	0.98	0.14	0.19	0.05	0.17	0.13	0.98
	10 FIX	*t*	10.5(4.3)	1.2(0.4)	2.4(1.1)	1.3(0.5)	0.60	0.65	0.89	0.94	0.76	0.80	0.02	0.13	0.60	0.94
	15 FIX	*t*	10.7(2.9)	1.7(0.5)	1.0(0.7)	2.0(0.5)	0.15	0.25	0.88	0.98	0.15	0.21	0.06	0.21	0.15	0.98
	INCR	t,t^2	9.8(3.7)	1.6(0.5)	0.9(0.6)	1.3(0.6)	0.20	0.29	1.00	1.00	0.08	0.10	0.00	0.01	0.20	1.00
	10 FIX	t,t^2	11.4(3.6)	1.1(0.3)	2.8(1.3)	1.5(1.0)	0.64	0.68	0.89	0.93	0.82	0.83	0.05	0.13	0.64	0.93
	15 FIX	t,t^2	10.9(2.9)	1.7(0.5)	0.8(0.5)	1.3(0.6)	0.22	0.28	1.00	1.00	0.03	0.04	0.00	0.01	0.22	1.00
m2,y,p,i	INCR	*t*	7.4(2.5)	1.7(0.5)	0.9(1.0)	2.6(1.0)	0.88	0.95	1.00	1.00	0.60	0.63	0.38	0.45	0.88	1.00
	10 FIX	*t*	12.2(2.7)	1.6(0.5)	2.7(0.8)	2.0(1.0)	0.74	0.79	0.26	0.30	0.87	0.89	0.01	0.05	0.74	0.30
	15 FIX	*t*	7.8(2.1)	1.8(0.4)	0.7(0.9)	2.7(0.9)	0.95	0.98	1.00	1.00	0.70	0.75	0.47	0.49	0.95	1.00
	INCR	t,t^2	7.2(4.0)	1.7(0.5)	1.5(1.5)	2.3(1.2)	0.87	0.94	1.00	1.00	0.39	0.46	0.02	0.03	0.87	1.00
	10 FIX	t,t^2	12.4(2.3)	1.6(0.5)	3.0(1.2)	1.3(0.9)	0.74	0.79	0.26	0.36	0.89	0.90	0.08	0.15	0.74	0.36
	15 FIX	t,t^2	7.7(4.0)	1.8(0.4)	1.6(1.6)	2.5(1.2)	0.95	1.00	1.00	1.00	0.45	0.54	0.02	0.04	0.95	1.00
m1,y,p,i,c	INCR	*t*	10.0(3.4)	1.6(0.5)	2.1(1.0)	2.8(0.7)	0.14	0.18	0.91	1.00	0.27	0.36	0.13	0.21	0.14	1.00
	10 FIX	*t*	13.0(0.2)	1.0(0.1)	4.1(0.8)	1.5(0.6)	0.69	0.79	1.00	1.00	0.91	0.94	0.03	0.12	0.69	1.00
	15 FIX	*t*	10.1(3.0)	1.7(0.5)	1.8(0.6)	2.9(0.7)	0.13	0.16	0.91	1.00	0.19	0.29	0.15	0.26	0.13	1.00
	INCR	t,t^2	10.1(3.4)	1.6(0.5)	2.0(0.7)	2.0(0.8)	0.17	0.28	1.00	1.00	0.19	0.29	0.05	0.15	0.17	1.00
	10 FIX	t,t^2	13.0(0.1)	1.0(0.0)	4.3(0.9)	1.7(0.8)	0.77	0.83	0.00	0.00	0.96	0.98	0.09	0.16	0.77	0.00
	15 FIX	t,t^2	10.1(3.0)	1.7(0.5)	1.9(0.4)	2.1(0.8)	0.17	0.25	1.00	1.00	0.07	0.19	0.06	0.19	0.17	1.00
m2,y,p,i,c	INCR	*t*	10.1(3.0)	1.7(0.45)	2.0(1.0)	3.2(0.8)	0.66	0.83	1.00	1.00	0.41	0.42	0.25	0.29	0.66	1.00
	10 FIX	*t*	13.0(0.00)	1.3(0.44)	4.3(0.7)	2.5(1.0)	0.74	0.78	0.26	0.26	0.95	0.97	0.02	0.06	0.74	0.26
	15 FIX	*t*	9.4(2.9)	1.9(0.33)	1.7(0.6)	3.3(0.6)	0.65	0.80	1.00	1.00	0.28	0.29	0.26	0.27	0.65	1.00
	INCR	t,t^2	10.1(3.0)	1.7(0.46)	1.9(1.0)	2.9(1.3)	0.65	0.73	1.00	1.00	0.19	0.21	0.00	0.00	0.65	1.00
	10 FIX	t,t^2	13.0(0.00)	1.1(0.28)	4.2(0.8)	2.0(0.8)	0.75	0.79	0.43	0.43	0.99	1.00	0.00	0.01	0.75	0.43
	15 FIX	t,t^2	9.5(3.0)	1.9(0.36)	1.6(0.6)	3.1(1.3)	0.63	0.68	1.00	1.00	0.02	0.03	0.00	0.00	0.63	1.00

[a]The table summarizes results for rolling estimations based on monthly data from 1960:2–1996:3. The variables used in each VAR and VEC model are recorded under the heading 'Model'. Estimation of the models is based on rolling 10 and 15 year fixed windows of data (10 FIX and 15 FIX) as well as 10 year increasing windows (INCR). The averages and standard errors of estimated lag orders and estimated cointegrating ranks across all sub-samples are given in the first four columns of numerical entries in the table. The third column indicates whether the data are linearly (t) or quadratically (t and t^2) detrended, and applies to the levels data. Empirical Granger causality test rejection frequencies at the 5% and 10% level are given in the 8th to 15th columns. These values correspond to our 'classical' difference and surplus lag Wald tests. Values are reported for models which are estimated using lag orders selected by both the AIC and SIC. The final 2 columns in the table correspond to model selection based predictive accuracy tests. Using the model selection approach discussed above, the lag order of the models is first selected based on either the AIC or the SIC, and then competing forecasting models are compared, and the 'best' is selected, also based on either the AIC or the SIC, with the same criterion used to select lags and choose the 'best' model. Numerical values indicate the proportion of times that the 'best' model contains money.

an even more severe degree of freedom problem than was the case for the difference Wald tests), we must tentatively conclude that there is little evidence of the marginal predictive content of money for income, based on surplus lag tests. However, these tests may suffer from reduced finite sample power. Furthermore, we know that our difference VAR Wald tests are subject to finite sample size bias. Indeed, it is quite likely that these two related, but opposite problems account in large part for the disparity in our findings based on the two classical Granger noncausality tests. This is one of the reasons why we use a model selection approach in addition to Granger causality tests.

As outlined above, our model selection based predictive accuracy tests are based on the premise of choosing the 'best' forecasting model. For this reason, we might wish to entertain the rather widely held notion that more parsimonious models (i.e. those specified using the SIC to select lag order) often forecast better. In order to evaluate this hypothesis, we carried out a series of simple experiments. Using all of our different windows of data, we fit levels VAR models, difference VAR models, and VEC models to 0.7^*T of each rolling sample of data. The remaining 0.3^*T observations in each rolling sample were used to construct a sequence a rolling one-step ahead ex ante forecasts of income. A total of 0.3^*T new models, cointegrating spaces, etc. were estimated for each roll of the data window, in order to construct the 0.3^*T length one-step ahead forecast sequences (so that many thousands of models were estimated in all). In all cases, competing models were estimated which used the AIC and the SIC to select lag order, and Diebold and Mariano (1995) loss differential tests statistics were constructed for each window of data (based on mean square forecast error, mean absolute forecast error deviation, and mean absolute percentage forecast error – see Swanson and White (1997) for details of these statistics) to determine whether one lag selection procedure resulted in better forecast models, in the sense of quadratic forecast error loss. The experiment was repeated for both trend specifications. As might be expected, the results based on these ex ante experiments suggest that the SIC is preferred by a more than a 2 to 1 ratio, regardless of model specification. Summarizing our findings, the AIC tends to pick the best out-of-sample model around 0–20% of the time, the SIC around 30–65% of the time, and neither model is preferred around 30–50% of the time. (To save on space, tabulated results are not included.) Thus, we have direct evidence that the SIC is preferable when constructing one-step ahead forecasting models of income. For this reason, we prefer predictive accuracy tests which use models estimated with lag orders chosen using the SIC.

The results based on our model selection approach are gathered under the heading 'predictive accuracy tests' in Table 1. Note that for the SIC version of this test, both lag order and the 'best' forecasting model are selected using the SIC. Analogously, the AIC version of the test uses the AIC for lag selection and for choosing the 'best' forecasting model. Interestingly, using the SIC results in the selection of models which include money close to 100% of the time,

regardless of variables, trend specification, window, etc. The exception to this result is the 10 year fixed window of data, which again appears to be too short to offer relevant evidence in our experiments, at least when $m2_t$ is used. Since the SIC can be interpreted as a statistic useful for selecting optimal forecasting models, we have rather strong new evidence of the marginal predictive content of money for income. This evidence comes to light in part by examining carefully the shortcomings of other tests of Granger noncausality in the presence of cointegrated data, and is clearly robust to sample period, as long as the data horizon used is sufficiently long. Further, as the VEC models estimated for our predictive accuracy tests generally include a cointegrating vector which is the c_t-i_t spread, we have some evidence that Friedman and Kuttner's (1993) preferred model (see above discussion) actually supports a finding of money income causality, when viewed through a rolling window. Also, estimation of Thoma's (1995) preferred model (which is essentially a difference VAR model, and can be examined using the difference Wald test) leads to strong evidence of the marginal predictive content of money for income, subject to the above test bias criticism, of course. Finally, we come back to Stock and Watson's (1989) original findings based on systems such as those examined here. Our evidence, which is based on a somewhat broader examination of many related models, clearly supports their original finding of the predictive ability of money for income.

6. Summary and concluding remarks

We have used a rolling window approach based on fixed and increasing samples of data to investigate the extent to which fluctuations in the money stock anticipate fluctuations in real output. Based on our empirical analysis, we offer the following conclusions. First, it is far from clear that money does not have significant predictive power for income. When either Divisia monetary aggregates or simple sum money measures are used as the monetary aggregate in systems with income, prices, and interest rates, the predictive content of money is significant for virtually all of the rolling samples between 1960:2 and 1994:6. However, this result is somewhat dependent upon the type of test which is used. For example, predictive accuracy tests which are based on a model selection approach to selecting 'best' forecasting models are shown to overwhelmingly favour models with money, even when additional regressors, such as the commercial paper – T-bill rate spread are added to our models. Other hypothesis tests, while less favorable to money, are shown to be fraught with potential problems. In particular, we note that Wald tests based on differenced data VAR models may be biased when the true data generating process is a VEC model. Surplus lag Wald tests, on the other hand, while accounting for cointegration of unspecified form, may be affected by reduced finite sample power.

Second, the systems of variables examined appear to evolve to some final form over time, in particular with respect to the characterization of the basis of the cointegration space. More precisely, we find no advantage to considering shorter 10 year fixed windows of data. The systems are particularly stable when the SIC is used to select the lag order. Furthermore, the SIC is preferred to the AIC, when used to select the lag order of VEC models used to produce one-step ahead ex ante forecasts of income.

The work here is merely a starting point. A wide variety of further questions present themselves for subsequent research, both theoretical and empirical. On the theoretical side, it is of interest to establish the statistical properties of surplus lag Wald tests, particularly with respect to test power. Interesting empirical projects include applying the analysis in this paper to the examination of impulse response functions, investigating the predictive power of money for income using alternative measures of predictive ability (such as rolling window forecast error-variance decompositions in a vector error-correction system), constructing true ex ante model selection based tests for the predictive ability of money, and expanding the endogenous systems analyzed by including labor demand channels (as measured by the unemployment rate), for example.

Acknowledgements

Many thanks to Shaghil Ahmed, Lawrence Christiano, R.W. Hafer, Ping Wang, Halbert White, and to seminar participants at the University of Toronto and Queen's University for useful comments and discussions. Thanks also to Robert King and an anonymous referee for detailed and extensive comments and suggestions. Financial support from the Research and Graduate Studies Office at Pennsylvania State University is gratefully acknowledged.

References

Anderson, R.G., Jones, B., Nesmith, T., 1996. Building new monetary services indices: concepts, methodology and data. Working paper, Federal Reserve Bank of St. Louis, St. Louis, MI.

Barnett, W.A., 1978. The user cost of money. Economic Letters 1, 145–149.

Barnett, W.A., 1980. Economic monetary aggregates: an application of index number and aggregation theory. Journal of Econometrics 14, 11–48.

Barnett, W.A., 1990. Developments in monetary aggregation theory. Journal of Policy Modeling 12, 205–257.

Bernanke, B.S., 1986. Alternative explanations of the money-income correlation. Carnegie–Rochester Conference Series on Public Policy 25, 49–100.

Bewley, R., Orden, D., Xang, M., Fisher, L.A., 1994. Comparison of Box–Tiao and Johansen canonical estimators of cointegrating vector in VEC(1) models. Journal of Econometrics 64, 3–27.

Blanchard, O.J., Quah, D., 1989. The dynamic effects of aggregate demand and supply disturbances. American Economic Review 79, 655–673.

Christiano, L.J., Ljungqvist, L., 1988. Money does Granger-cause output in the bivariate money–output relation. Journal of Monetary Economics 22, 217–235.
Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–263.
Dolado, J.J., Lütkepohl, H., 1996. Making Wald tests work for cointegrated VAR systems. Econometric Reviews 15, 369–386.
Friedman, B.M., Kuttner, K.N., 1993. Another look at the evidence on money-income causality. Journal of Econometrics 57, 189–203.
Gonzalo, J., 1994. Five alternative methods of estimating long-run equilibrium relationships. Journal of Econometrics 60, 203–233.
Granger, C.W.J., 1996. Can we improve the perceived quality of economic forecasts? Journal of Applied Econometrics 11, 455–473.
Granger, C.W.J., King, M.L., White, H., 1995. Comments on testing economic theories and the use of model selection criteria. Journal of Econometrics 67, 173–187.
Granger, C.W.J., Lin, J.-L., 1995. Causality in the long run. Econometric Theory 11, 530–536.
Granger, C.W.J., Teräsvirta, T., 1993. Modeling Nonlinear Economic Relationships. Oxford, New York, NY.
Hafer, R.W., Jansen, D.W., 1991. The demand for money in the United States: evidence from cointegration tests. Journal of Money, Credit, and Banking 23, 155–168.
Hafer, R.W., Kutan, A.M., 1997. More evidence on the money-output relationship. Economic Inquiry 35, 48–58.
Johansen, S., 1988. Statistical analysis of cointegrating vectors. Journal of Economic Dynamics and Control 12, 231–254.
Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 59, 1551–1580.
King, R.G., Plosser, C.I., Stock, J.H., Watson, M.M., 1991. Stochastic trends and economic fluctuations. American Economic Review 81, 819–840.
King, R.G., Watson, M.M., 1992. Testing long run neutrality. Working paper, National Bureau of Economic Research, Boston, MA.
McCallum, B.T., 1979. The current state of the policy-ineffectiveness debate. American Economic Review 69, 240–245.
Potter, S., 1995. A nonlinear approach to US GNP. Journal of Applied Econometrics 10, 109–125.
Sims, C.A., 1980. Comparison of interwar and postwar cycles: monetarism reconsidered. American Economic Review 70, 250–257.
Stock, J.H., Watson, M.M., 1989. Interpreting the evidence on money-income causality. Journal of Econometrics 40, 161–181.
Stock, J.H., Watson, M.M., 1993. A simple estimator of cointegrating vectors in higher order integrated systems. Econometrica 61, 783–820.
Swanson, N.R., Ozyildirim, A., Pisu, M., 1996. A comparison of alternative causality and predictive accuracy tests in the presence of integrated and cointegrated economic variables. Working paper, Pennsylvania State University, University Park, PA.
Swanson, N.R., White, H., 1995. A model selection approach to assessing the information in the term structure using linear models and artificial neural networks. Journal of Business and Economic Statistics 13, 265–275.
Swanson, N.R., White, H., 1997. A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. Review of Economics and Statistics 79, 540–550.
Thoma, M.A., 1994. Subsample instability and asymmetries in money-income causality. Journal of Econometrics 64, 279–306.
Toda Hiro, Y., Yamamoto, T., 1995. Statistical inference in vector autoregressions with possibly integrated processes. Journal of Econometrics 66, 225–250.