ELSEVIER

Journal of Empirical Finance 8 (2001) 639–668

Journal of EMPIRICAL FINANCE

www.elsevier.com/locate/econbase

Estimation of a rational expectations model of the term structure ☆

Angelo Melino *

University of Toronto, 150 St. George St. Toronto, Canada, M553G7

Abstract

This paper shows how to represent a vector autoregression (VAR) in terms of the eigenvalues and eigenvectors of its companion matrix. This representation is used to impose the exact restrictions implied by the expectations hypothesis on the VAR for short and long term interest rates and to calculate the restricted maximum likelihood estimates. The first difference representation for short and long rates used by Sargent (1979) is shown to be inconsistent with the expectations hypothesis, but a VAR with two unit roots is constructed that satisfies the exact restrictions and leads to similar restricted estimates. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Rational expectations; Term structure; Restrictions

1. Introduction

In a recent paper, Sargent (1979) describes a novel method for estimating and testing an expectations model of the term structure of interest rates that is consistent with the rational expectations hypothesis. Sargent assumed that the first differences of the short and long rate followed a bivariate weakly stationary process. He then deduced the restrictions on the autocovariances of this bivariate process implied by a model which claims that agents act in such a way as to

☆ This was my "job-market" paper. A fairly complete draft was presented at various institutions in 1980–1981. This version of the paper, except for a few modest changes, appears as the first essay of my thesis [Melino, A., 1983. Essays on Estimation and Interference in Linear Rational Expectations Models. PhD dissertation, Harvard University].

* Fax: +1-416-978-6713.

E-mail address: melino@chass.utoronto.ca (A. Melino).

0927-5398/01/$ - see front matter
PII: S0927-5398(01)00041-X

equate the yield to maturity of a long bond to the average of optimally forecasted future short rates.

The assumption that agents use optimal linear forecasts imposes fairly complicated nonlinear restrictions on the covariance structure of the yields on short and long-term bonds. Sargent was able to test these restrictions and concluded that they were not significantly violated by the particular data set that he examined.

However, Sargent did not impose all of the covariance restrictions implied by the model. Basically, he tested whether the error in forecasting the difference in the 3-month holding period yield of short and long bonds was uncorrelated with lagged values of the short and long rates. As usually formulated, however, the model also implies that this difference should also be uncorrelated with the *current* values of the short and long rate.

These last two orthogonality restrictions are particularly interesting,[1] and one of the main motivations for this paper is to examine the implications of imposing them.

One of the paper's results is that if we impose these extra restrictions, we can reject the model proposed by Sargent with probability one. However, this result must be interpreted carefully. Because of certain technical reasons which we shall elaborate later, it turns out that it is the first difference representation assumed by Sargent which is basically incompatible with the model. By relaxing this representation, we can estimate a model very similar in spirit and find that Sargent's original conclusions still hold.

One of the distinguishing features of this paper is that we exploit results described elsewhere (Melino, 1983, Chap. 2) that allow us to obtain a closed form representation of the rational expectations restrictions. The basic idea is to parameterize the model explicitly in terms of the roots of the vector autoregression, and, if necessary, certain other parameters.

This alternative parameterization has several attractive features. From a theoretical point of view, the analytic representation that it provides is useful for deriving various properties of the restrictions and for improving our understanding of their implications. For example, it turns out that there are quite a number of different ways in which the term structure restrictions can be satisfied. While it is hard to provide any economic interpretation to these various alternatives, the approach we have chosen at least allows us to describe them, and provides a feasible computational method for investigating each alternative in turn.

The plan of this paper is as follows. In Section 2, we review the model and Sargent's proposed estimation scheme. In Section 3, we outline our proposed parameterization and discuss some of its features. Section 4 reports the results of

[1] One can show that these particular restrictions are the basis for the volatility tests recently proposed by Shiller (1979). Shiller concluded that there is ample evidence to reject the expectations model.

estimating Sargent's model while imposing all of the relevant restrictions. In Section 5, we attack the estimation problem using standard large sample theory. The approach suggested by Wald (1943) seems particularly applicable to linear rational expectations models. For some reason, however, it has not been exploited by researchers. In our application, it can be reported that Wald's test and estimator performed very well. Section 6 deals briefly with some possible extensions which we hope to pursue in future research. Finally, following a well established tradition, we end with a brief conclusion.

2. The model

Expectations models of the term structure of interest rates have a long history in the economics literature. Since the primary focus of this paper is on the nuts and bolts of estimation, no attempt to motivate the model will be made here, and instead the reader is referred to the paper by Sargent (1979) and the references therein.

Let r_t and R_t denote, respectively, the yields to maturity on a short and long bond. We begin by assuming that the vector $\boldsymbol{y}_t = (r_t R_t)'$ is drawn from a stochastic process with a representation of the form[2]

$$\boldsymbol{y}_t = \sum_{i=1}^{L} \mathbf{a}_i y_{t-i} + \boldsymbol{e}_t \tag{2.1}$$

where $\mathbf{a}_i$ are (2×2) matrices of coefficients of the form

$$\mathbf{a}_i = \begin{vmatrix} \alpha_i & \beta_i \\ \gamma_i & \delta_i \end{vmatrix} \tag{2.2}$$

and $\boldsymbol{e}_t$ is a vector white noise process with contemporaneous covariance matrix $\boldsymbol{\Sigma}$.[3]

Eq. (2.1) can be written compactly as

$$\boldsymbol{Y}_t = \mathbf{A}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t \tag{2.3}$$

where $\boldsymbol{Y}_t$ and $\boldsymbol{E}_t$ are $(2L \times 1)$ vectors of the form

$$\boldsymbol{Y}_t = (\boldsymbol{y}_t \boldsymbol{y}_{t-1} \cdots y_{t-L+1})'$$
$$\boldsymbol{E}_t = (\boldsymbol{e}_t 0 \ldots 0)' \tag{2.3a}$$

[2] The intercepts in Eq. (2.1) have been suppressed so that the data should be thought of as deviations from the sample means.

[3] In practice, the lag length L is chosen large enough for the fitted $\boldsymbol{e}_t$ to look like white noise.

and **A** is a $(2L \times 2L)$ matrix given by

$$\begin{vmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_{L-1} & \boldsymbol{a}_L \\ \boldsymbol{I} & 0 & \cdots & 0 & 0 \\ 0 & \boldsymbol{I} & \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & \cdots & \boldsymbol{I} & 0 \end{vmatrix} \tag{2.3b}$$

The representation (2.3), which allows us to express a high order distributed lag as a first-order Markov process, is sometimes referred to as the "companion form," and the matrix **A** is called the "companion matrix."[4]

It is well known that an autoregressive representation of the above form provides a suitable approximation for almost all covariance stationary processes. Moreover, certain mild kinds of nonstationarity such as random walks or explosive systems are not ruled out by Eq. (2.1).

Nonetheless, the assumption that a representation of the form as Eq. (2.1) exists is not innocuous. Proceeding formally, we can in each period represent $(r_t R_t)'$ as the sum of a projection on lagged values and its orthogonal complement. However, there may be no good reason to believe that the coefficients of the projection should remain constant over time. Lucas (1976) points to the instability of policy rules as a major cause of shifting behavioural equations and, hence, sample covariances. At an empirical level, the fact that regression coefficients are sensitive to the sample period is well known.

It should be noted, though, that for the purposes at hand, we can proceed under somewhat weaker assumptions. These will be discussed briefly below. In the final analysis, however, we cannot avoid the requirement that some stable relationship exist between future and lagged values of the data.

2.1. The restrictions

One version of the rational expectations (RE) model of the term structure can be written as[5]

$$R_t = \frac{1}{N} \sum_{i=0}^{N-1} E(r_{t+i} \mid I_t) \tag{2.4}$$

[4] Gantmatcher (1959), p. 149.

[5] Shiller (1979) provides a useful list of various expectations models, and a common framework for deriving them. Notice that since the data are treated as deviations from means (see footnote 2), Eq. (2.4) implicitly allows for a nonzero time invariant term premium.

where N is the maturity of the long bond, and $E(X_{t+s} \mid I_t)$ denotes the least squares projection[6] of X_{t+s} on the information set I_t available to agents at time t. I_t is assumed to include at least all current and lagged values of the short and long rates.

Equations similar to Eq. (2.4), which relate one variable to a weighted average of optimal forecasts of future values of a second variable, are common in economics (e.g., permanent income consumption function) and arise naturally as solutions to intertemporal optimization problems. Eq. (2.4) is notably different in that it is assumed to hold without error. Usually, some sort of error is tacked on to take care of possible errors of specification, or to capture the notion that the restriction only holds "approximately." The somewhat ad hoc addition of an error term raises a host of econometric and interpretive problems. Nonetheless, by not including such an error term, the model has certain implications that are probably stronger than the intuitive notion which Eq. (2.4) is supposed to capture.[7]

Since the information set available to agents is very large, there is no possibility of testing the specification given by Eq. (2.4) directly. In order to proceed, we must relate the RHS of Eq. (2.4) to an expression that is estimable. Let I_t' denote any information set contained in I_t. Then, using the law of iterated projections, it is easy to show that

$$E(R_t \mid I_t') = \frac{1}{N}\sum_{i=0}^{N-1} E\big(E(r_{t+i} \mid I_t) \mid I_t'\big) = \frac{1}{N}\sum_{i=0}^{N-1} E(r_{t+i} \mid I_t') \tag{2.5}$$

For a suitably chosen set I_t', we can, in principle at least, estimate both sides of Eq. (2.5) and test if they are equal. However, the choice of I_t' is somewhat problematic. From the viewpoint of the null hypothesis, of course, any set contained in I_t will do. However, in order to have a good chance of rejecting the model, we want to include in I_t' those variables, which are likely to affect the left and right hand sides of Eq. (2.5) differently. For example, if we believe that in the real world, the markets for short- and long-term debt instruments are largely segmented, we would like to include in I_t' variables which are fairly specific to one of the two markets. In some models, for example, wealth flows to pension funds or insurance companies may be important in explaining short-term movements in the long rate and yet have little effect on the sequence of future short rates.

From a conceptual point of view, there is no problem in using the procedure described below to deal with any arbitrary information set. It is convenient, though, to keep the dimension of I_t' small. As a first step, therefore, we shall

[6] One could begin with the stronger assumption that Eq. (2.4) holds with $E(X_{t+s} \mid I_t)$ denoting the conditional expectation of X_{t+s}. In this paper, though, the distinction has no operational content. Since we are dealing only with second moment properties of the process, it seems preferable to use the somewhat less restrictive terminology.

[7] Note that adding a residual to the RHS of Eq. (2.4) invalidates Shiller's volatility test.

follow Sargent's lead and take I_t' to be the current and lagged values of the short and long rates. Note that this choice implies that the LHS of Eq. (2.5) is identically equal to R_t, as $R_t \in I_t'$.

It will prove convenient to express Eq. (2.5) as restrictions on the matrix **A** given by Eq. (2.3). First note that $E(r_{t+i} \mid I_t')$ is just the first element in the vector $E(\mathbf{Y}_{t+i} \mid I_t')$. Further from Eq. (2.3), we deduce that $E(\mathbf{Y}_{t+i} \mid I_t') = \mathbf{A}^i \mathbf{Y}_t$.[8] Therefore, we can rewrite Eq. (2.5) as

$$e_2' \mathbf{Y}_t = \frac{1}{N} e_1' (\mathbf{I} + \mathbf{A} + \ldots + \mathbf{A}^{N-1}) \mathbf{Y}_t \tag{2.6}$$

where e_j is a $(2L \times 1)$ vector with unity in the jth position and 0 elsewhere.

Eq. (2.6) is postulated to hold at all times, regardless of the realized values of the short and long rates. Unless $\mathbf{Y}_t$ is a singular process,[9] the only way that Eq. (2.6) can be satisfied is to require that the nontrivial rows of **A** satisfy restrictions of the form

$$e_2' \mathbf{I} = \frac{1}{N} e_1' (\mathbf{I} + \mathbf{A} + \ldots + \mathbf{A}^{N-1}) \tag{2.7}$$

Eq. (2.7) provides a concise summary of the restrictions implied by both rational expectations and the arbitrage relationship described in Eq. (2.4). In the absence of any further restrictions on the matrix **A**, it can be shown that solutions of Eq. (2.7) exist.[10] However, it may not be possible to satisfy both Eq. (2.7) and other a priori restrictions.[11]

For example, suppose that we believe the short rate follows some AR(L) process, with L finite, that is independent of the long rate process, i.e., $\beta = 0$. It is not hard to see that Eq. (2.7) becomes inconsistent. The interpretation to make is not that the RE restrictions cannot be satisfied, but that coupled with the assumption $\beta = 0$, they imply that $\mathbf{Y}_t$ must be a singular process. This, of course, is a very strong implication that is most unlikely to be satisfied by real world data sets.

Except for the above proviso, it should be stressed that Eq. (2.7) exhausts the second-order restrictions on the short and long rate process. In particular, the appropriate analogue of the "volatility" restrictions of Shiller (1979) and Singleton (1980) are included in the implications of Eq. (2.7), and there are no further implied restrictions on the (conditional) contemporaneous covariance matrix $\mathbf{\Sigma}$.

[8] We rely heavily on the fact that L has been chosen in Eq. (2.1) large enough to make $\mathbf{e}_t$ look like white noise. This allows us to deduce that $E(\mathbf{Y}_{t+i} \mid \mathbf{Y}_t, \mathbf{Y}_{t-i} \ldots) = E(\mathbf{Y}_{t+i} \mid \mathbf{Y}_t)$.

[9] We shall say that the $\mathbf{Y}_t$ process is singular if $E(\mathbf{Y}_t \mathbf{Y}_t')$ has less than full rank. If this occurs, then we can find a vector of constants, say $\mathbf{c}$, not all zero such that $\mathbf{c}' \mathbf{Y}_t = 0$ with probability one. If the first component of $\mathbf{c}$ does not equal 0, we can interpret this as meaning that the short rate can be predicted perfectly just with knowledge of the current long rate, and lagged values of the short and long rates.

[10] At least in the nontrivial case $N > 1$. If $N = 1$, the model amounts to the law of one price, $r_t = R_t$ for all t.

[11] See also the discussion in Section 4.

2.2. Sargent's estimation procedure

One way of imposing the RE restrictions is to view Eq. (2.7) as defining an implicit relationship across the nontrivial rows of **A**. Formally, we can write $(\gamma, \delta) = \phi(\alpha, \beta)$. Instead of maximizing the likelihood function $L(\alpha, \beta, \gamma, \delta, \mathbf{\Sigma})$, the restricted problem has us maximize $L(\alpha, \beta, \phi(\alpha,\beta), \mathbf{\Sigma})$.[12]

The problem would be entirely straightforward and standard if $\phi(\cdot)$ was easy to evaluate. In practice, however, solving for the values of (γ, δ) consistent with a set of given values of (α, β) can be quite difficult.

Sargent proposed the following iterative procedure for solving these restrictions. Construct the matrix $\mathbf{A}(0)$, of the form described in Eq. (2.3b), with its first row set equal to the given values of (α, β) and second row set equal to some initial guess of $\phi(\alpha, \beta)$, say $(\gamma(0), \delta(0))$. We can construct a new estimate by forming

$$(\gamma(1),\delta(1)) = e_2' \mathbf{A}(1) = \frac{1}{N} e_1' \left(\mathbf{A}(0) + \mathbf{A}^2(0) + \ldots \mathbf{A}^N(0)\right) \tag{2.8}$$

If we repeat this process, keeping all but the second row fixed, until $(\gamma(n), \delta(n))$ converges, then in most cases we have found a solution of Eq. (2.7).

In general, Eqs. (2.7) and (2.8) are not equivalent. Eq. (2.7) follows from assuming that I_t includes both current and lagged values of the short and long rate. By contrast, Eq. (2.8) holds even if I_t does not contain the current values of $(r_t R_t)'$.[13] Mathematically, Eq. (2.8) is obtained by postmultiplying the system Eq. (2.7) by the matrix **A**. If **A** is nonsingular, then we can retrieve all of the original restrictions from Eq. (2.8). However, as we shall see in Section 3, for certain singular **A** matrices, the rank of the restrictions in Eq. (2.8) is less than that implied by Eq. (2.7). For practical purposes, the importance of the result is that by using Sargent's proposed algorithm, we may be unwittingly giving up some of the testable implications of the original expectations model.

While numerical methods are often satisfactory, there are good reasons for preferring a closed form solution. The latter allows us to identify and hopefully avoid certain pitfalls.[14] More importantly, it can contribute to a better understanding of the model.

---

[12] In the absence of restrictions on $\mathbf{\Sigma}$, we can concentrate the likelihood function so that maximizing $L(\cdot)$ is equivalent to minimizing $\det(\hat{\mathbf{\Sigma}})$.

[13] This difference was not a problem in Sargent's paper, since he dealt only with projections on information available at time $t-1$.

[14] Sargent reported that this algorithm worked well in his application. Also, it seemed to work best when the roots of the autoregression were well inside the unit circle and the initial estimate of $(\gamma(0), \delta(0))$ was set equal to zero. (See Section 4 for a discussion of this latter condition.) Since Sargent was dealing with first differenced data, the former condition was usually satisfied. In my experience, however, using the levels of the data, Sargent's algorithm worked very poorly.

3. An alternative parameterization

In this section we introduce an alternative parameterization of the vector autoregression and show how it can be used to obtain a closed form representation of the rational expectations restrictions given in Eq. (2.7).

We begin with the observation that any matrix **A** can be written in the form[15]

$$\mathbf{A} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1} \tag{3.1}$$

where $\boldsymbol{\Lambda}$ is a $(2L \times 2L)$ matrix given by $\boldsymbol{\Lambda} = \text{diag}\,(\boldsymbol{\Lambda}_{k_i}(\lambda_i))$; $\boldsymbol{\Lambda}_1(\lambda) = \lambda$, and if $k_i > 1$,

$$\boldsymbol{\Lambda}_{k_i}(\lambda_i) = \begin{vmatrix} \lambda_i & 1 & 0 & \cdot & \cdot & 0 \\ 0 & \lambda_i & 1 & \cdot & \cdot & 0 \\ 0 & 0 & \lambda_i & \cdot & \cdot & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & \cdot & \cdot & 1 \\ 0 & 0 & 0 & \cdot & \cdot & \lambda_i \end{vmatrix}$$

is a $(k_i \times k_i)$ matrix with one of the eigenvalues of **A**, λ_i along the main diagonal, unity along the superdiagonal and zeros elsewhere, and **P** is a $(2L \times 2L)$ nonsingular matrix whose columns are the corresponding generalized eigenvectors.[16] The representation given by Eq. (3.1) is often called the Jordan (canonical) form of the matrix **A**.

The Jordan form is particularly useful in evaluating matrix polynomials so that it may be easier to think about the restrictions in Eq. (2.7) if we reason directly from the decomposition afforded by Eq. (3.1). Since there is a one to one association between the Jordan decomposition and the matrix **A**,[17] there is no reason to feel committed to either representation. We are free, of course, to choose whichever parameterization is most convenient for the question at hand.

The features of the matrix **A** in Eq. (2.3b) are reflected in restrictions on the form that the corresponding **P** and $\boldsymbol{\Lambda}$ matrices can take. If we choose to specify the vector autoregression directly from the Jordan decomposition, we must impose these restrictions a priori in order to make it an equivalent parameterization. It is convenient to think of these restrictions as being of three basic types.

[15] Most of the properties of matrices used in this section are discussed in standard textbooks. Gantmacher (1959) is an advanced, but particularly useful reference.

[16] The columns of **P** corresponding to $\boldsymbol{\Lambda}_i$ are also called a Jordan chain.

[17] Given one of the normalization rules discussed below.

First of all, notice that **P** can always be replaced by the matrix **PK**, where **K** is any nonsingular matrix that commutes with $\mathbf{\Lambda}$.[18] This inherent indeterminancy allows us to impose certain normalization rules, which in effect reduce the number of free parameters in **P**.

Secondly, since the elements of **P** and $\mathbf{\Lambda}$ are to be thought of as complex numbers, we must require that they satisfy certain restrictions in order to ensure that the resulting **A** matrix is real valued.

Finally, we are only interested in the Jordan form of matrices which have the special structure of the matrix **A** given by Eq. (2.3b). This requirement can be thought of as a third set of restrictions.

Imposing these restrictions turns out to be surprisingly simple, and reduces the number of free parameters needed to construct the matrices $\mathbf{\Lambda}$ and **P** to the number of unrestricted elements in **A**. For expositional purposes, we will first consider the case where **A** is known a priori to be of simple structure.[19] We will then deal briefly with the complications, which arise in the general case.

It turns out that the special structure of the matrix **A** in Eq. (2.3) is reflected in very simple restrictions on the columns of **P**. Let $\boldsymbol{P}_i$ be the ith eigenvector of **A**. Call its last two elements $\boldsymbol{p}_i = (p_{1i}, p_{2i})'$. Since the eigenvectors solve

$$\mathbf{A}\,\boldsymbol{P}_i = \lambda_i \boldsymbol{P}_i \tag{3.2}$$

we can use the equations 3 to $2L$ of the system (3.2) to deduce that the eigenvectors of **A** must be of the form

$$\boldsymbol{P}_i = \begin{vmatrix} \lambda_i^{L-1} & \boldsymbol{p}_i \\ \lambda_i^{L-2} & \boldsymbol{p}_i \\ \cdot & \\ \cdot & \\ \lambda_i & \boldsymbol{p}_i \\ & \boldsymbol{p}_i \end{vmatrix}$$

or

$$\boldsymbol{P}_i = \tilde{\boldsymbol{\lambda}}_i \otimes \boldsymbol{p}_i \tag{3.3}$$

[18] If $\mathbf{\Lambda}$ is diagonal, then so is **K**. In general, $\mathbf{K} = \text{diag}\,(\mathbf{K}_i)$, where $\mathbf{K}_i$ has the same dimensions as $\mathbf{\Lambda}_{k_i}$, and is an upper triangle matrix of the form

$$\mathbf{K}_i = \begin{vmatrix} k_{1i} & k_{2i} & \cdot & \cdot & \cdot \\ 0 & k_{1i} & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & k_{2i} \\ 0 & 0 & \cdot & \cdot & k_{1i} \end{vmatrix}$$

[19] A matrix is said to be of simple structure if it is similar to a diagonal matrix.

where $\tilde{\boldsymbol{\lambda}}_i = (\lambda_i^{L-1}, \lambda_i^{L-2} \ldots, \lambda_i, 1)'$, and " $\otimes$ " denotes the Kronecker product. A matrix whose columns are of the form given by Eq. (3.3) is called a Vandermonde matrix.[20]

Given the special structure of $\mathbf{P}$, it is easy to impose the remaining restrictions. The normalization rule allows us to express p_{1i} as a function of p_{2i}. There are various ways of doing this. A common technique is to choose $\boldsymbol{p}_i$ subject to $\boldsymbol{p}_i^* \boldsymbol{p}_i = 1$.[21] For simplicity, we may prefer to set $p_{1i} = 1$.[22]

Finally let C' denote the set of pairs (X, Y) such that X and Y are either both real numbers, or a complex conjugate pair. We satisfy the requirement that $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}$ is real valued by considering only pairs $((\lambda_i, \lambda_{i+1}), (p_{2i}, p_{2i+1})) \in C' \times C'$.

Note that we have reduced the number of free parameters needed to construct $\mathbf{P}$ and $\boldsymbol{\Lambda}$ to $4L$—the same as the number of nontrivial elements in the matrix $\mathbf{A}$. Rather than describe the matrix $\mathbf{A}$ by the vector $(\alpha, \beta, \gamma, \delta)$, we can describe it by specifying the appropriate values of $\boldsymbol{\lambda}$ and $\boldsymbol{p}$.

Armed with the knowledge that $\mathbf{A}$ is of simple structure, we could also imagine estimating it by maximizing the likelihood function over the permissible values of $\boldsymbol{\lambda}$ and $\boldsymbol{p}$. Of course, it would be silly to use this procedure to estimate the unrestricted model. However, when we impose the RE restrictions, this alternative becomes much more attractive.

If we replace the matrix $\mathbf{A}$ by $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}$ in Eq. (2.7), we see that we can express the restrictions as

$$e_2'\mathbf{P}\mathbf{P}^{-1} = \frac{1}{N} e_1'\mathbf{P} \left| \sum_{k=0}^{N-1} \boldsymbol{\Lambda} k \right| \mathbf{P}^{-1} \tag{3.4}$$

or just

$$e_2'\mathbf{P} = \frac{1}{N} e_1'\mathbf{P} \left| \sum_{k=0}^{N-1} \boldsymbol{\Lambda} k \right| \tag{3.4a}$$

Using the special structure of $\mathbf{P}$ given in Eq. (3.3), we see that this implies

$$\lambda_i^{L-1} p_{2i} = \lambda_i^{L-1} p_{1i} \frac{1}{N} \sum_{k=0}^{N-1} \lambda_i^k \qquad i = 1, \ldots 2L \tag{3.5}$$

Unless $\lambda_i = 0$,[23] we can impose the normalization rule $p_{1i} = 1$, and conclude that Eq. (2.7) using our parameterization reduces to

$$p_{2i} = S(\lambda_i) = \frac{1}{N} \sum_{k=0}^{N-1} \lambda_i^k \tag{3.6}$$

[20] The term is usually reserved for the univariate case, with $\boldsymbol{p}_i$ normalized to unity.
[21] The notation " * " is used to denote the conjugate transpose of a vector.
[22] This almost always works.
[23] If $L = 1$, Eq. (3.6) holds even for $\lambda_i = 0$.

The simplicity of Eq. (3.6) is refreshing, and is one of the main motivations for studying this particular parameterization. For nonsingular matrices, it is tempting to use Eq. (3.6) to express the restricted likelihood function just in terms of the eigenvalues of the matrix **A**.

It turns out to be inconvenient to search over pairs $(\lambda_i, \lambda_{i+1}) \in C'$ directly. In order to estimate the parameters, we would have to specify how many pairs are complex conjugate or real numbers, and then search over the permutations. This quickly becomes very tedious. However, any pair $(X, Y) \in C'$ can be thought of as the roots of a quadratic equation, say $z^2 + bz + c = 0$, with real coefficients (b, c). It seems simpler to specify the (b, c) pairs directly,[24] and then solve for the implied values of $(\lambda_i, \lambda_{i+1})$.

The likelihood function $\tilde{L}(b_i, c_i)$ can be evaluated as follows. We take the specified values of the (b, c) pairs and solve for the implied values of $\boldsymbol{\lambda}$. We can then use the normalization rule and Eq. (3.6) to construct the matrix **P**, calculate $\mathbf{P}^{-1}$, and then construct the matrix **A**. Once we have **A**, we evaluate the likelihood function at the implied values of $(\alpha, \beta, \gamma, \delta)$ in the usual way.

This may seem complicated, but it is actually fairly easy to implement on a computer. Complications do arise, however, when we extend the analysis to include matrices which are either singular or not of simple structure.

A serious difficulty with the Jordan representation is that it is really a list of various classes of matrices, most easily characterized by the structure of the $\boldsymbol{\Lambda}$ matrices. To be rigorous, one would have to consider the form of Eq. (2.7) for each of these structures separately, Unfortunately, the number of possible structures goes up very quickly with the dimension of **A**.

One way out is to appeal to an approximation theorem. A useful result is that we can approximate any matrix arbitrarily well by one having simple structure.[25] There are many circumstances where this "denseness" property can be invoked to argue that dealing only with matrices of simple structure imposes no real loss of generality. For our purposes, however, we need to know under what conditions a matrix satisfying Eq. (2.7) can be approximated arbitrarily well by a matrix of simple structure also satisfying Eq. (2.7). This is a fairly difficult problem, and I have not been able to solve it to my own satisfaction. However, experimentation suggests that there is no problem as long as the eigenvalue in $\boldsymbol{\Lambda}_i$ is nonzero.

Unfortunately, the case where the matrix **A** has multiple zero eigenvalues is substantially different, and in practice very important.

This might have been expected, since even in the simplest case, we see that there is a sharp difference between an eigenvalue that is exactly zero, and a "small" one. If λ_i is a very small nonzero number, then Eq. (3.6) tells us that p_{2i}

[24] If even, the characteristic polynomial can always be factored into quadratics with real coefficients. One can think of the (b, c) pairs as the coefficients in this factorization.

[25] See Bellman (1970), pp. 205–207.

must also be a small number. However, if $\lambda_i = 0$, then Eq. (3.6) is automatically satisfied and p_{2i} can take on any value.

The discontinuity of the restrictions around zero eigenvalues require us to deal with them explicitly. Unfortunately, this amounts to treating each possible case as a separate problem.

If we want to exploit the simple representation of the restrictions given by Eq. (3.6) in an estimation scheme, the tradeoff seems to be replacing one difficult problem with several simpler, but by no means trivial estimation problems.

The rules for parameterizing a singular matrix **A** are straightforward but tedious. The situation is complicated by the various forms that the Jordan matrix can take once we allow for multiple roots. Readers interested in the gory details are referred to Melino (1983, Chap. 2). For the purposes at hand, we need only note that up to a certain point, setting one of the eigenvalues equal to zero removes the restriction on the corresponding column of the Jordan chain. This introduces some flexibility in satisfying the restrictions given by Eq. (2.7), which turns out to be empirically important.

We can now summarize the contents of this section. In the case where **A** is nonsingular, we can take as the free parameters of the autoregression, the eigenvalue pairs of the matrix **A**. If **A** has zero roots of the characteristic polynomial, we can take as our free parameters the nonzero eigenvalues and the unrestricted parameters which appear in the generalized eigenvectors corresponding to the zero roots.[26] In practice, the likelihood function must be maximized over each of the cases $K = 0, 1 \ldots, L - 1$.[27]

Needless to say, the difficulty of dealing with multiple zero roots is a major drawback of estimating the autoregression directly from the Jordan decomposition. Even allowing for approximations, we still have to deal with at least as many estimation problems as there are lags of the autoregression.

3.1. Interpretation of the proposed parameterization

The purpose of this section is to provide some background material that may help to interpret the eigenvalues of **A**, and the parameters in **P**.

Consider once again the model

$$\boldsymbol{Y}_t = \mathbf{A}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t$$

Premultiplying this system of equations by the nonsingular matrix $\mathbf{P}^{-1}$ gives us

$$\mathbf{P}^{-1}\boldsymbol{Y}_t = \boldsymbol{\Lambda}\mathbf{P}^{-1}\boldsymbol{Y}_{t-1} + \mathbf{P}^{-1}\boldsymbol{E}_t$$

[26] While in general, the elements of **P** are complex numbers, it can be shown that the columns of **P** corresponding to a real valued eigenvalue are themselves real valued.

[27] We assume that the matrix **A** is nonderogatory. One can then show that the case $K \geq L$ is adequately approximated by the case $K = L - 1$ and all other roots distinct.

or

$$\mathbf{Z}_t = \mathbf{\Lambda}\,\mathbf{Z}_{t-1} + U_t \tag{3.10}$$

Eq. (3.10) is often referred to as the canonical representation of $\mathbf{Y}_t$. If all the roots of **A** are distinct, then Eq. (3.10) is nothing more than a vector of "seemingly unrelated" AR(1) processes, i.e., a vector of AR(1) processes with correlated residuals. The interpretation when **A** is not a simple matrix is similar, but not as straightforward. The eigenvalues of the matrix **A** have a simple interpretation in this particular coordinate system. Basically, they provide a measure of the persistence of shocks. If $|\lambda_i|$ is close to 0, then shocks to the $\mathbf{Z}_i$ process have only a very transitory effect. As $|\lambda_i|$ approaches unity, the shocks persist indefinitely. Finally, if $|\lambda_i|$ is greater than unity, shocks become amplified over time.

The interpretation of the elements in the matrix **P** is less straightforward. Consider once again the case where the matrix **A** has simple structure. Suppose we were given the values of $\mathbf{Z}_t$ and the matrix **P**. We could construct r_t by

$$r_t = \sum_{i=1}^{2L} p_{1i}\lambda_i^{L-1} Z_{it} \tag{3.11a}$$

Similarly, the long rate is given by

$$R_t = \sum_{i=1}^{2L} p_{2i}\lambda_i^{L-1} Z_{it} \tag{3.11b}$$

Given the normalization rule, we see that p_{1i} and p_{2i} measure the contribution of the $\mathbf{Z}_i$ process to the short rate relative to the long rate. The restrictions given by Eq. (3.6)

$$p_{2i}\lambda_i^{L-1} = p_{1i}\lambda_i^{L-1} S(\lambda_i)$$

therefore reflect the smoothing implication of the expectations models. For example, if the bivariate short and long rate process is covariance stationary, then all of the eigenvalues of **A** lie inside the complex unit circle, and the restrictions imply that each $\mathbf{Z}_i$ component contributes less to the long rate than to the short rate. In particular, therefore, in the covariance stationary case, the variance of the long rate must be less than the variance of the short rate.

3.2. The model in first differences

To test the algorithm suggested in Section 2.2, we employed it on the same data set as that used in Sargent (1979).[28] However, before estimating the model subject to all of the rational expectations restrictions, it seemed useful to reproduce

[28] The short rate is taken as the 3-month Treasury bill rate, whereas the long rate is the rate on 5-year government bonds. The sample period used for estimation is 1953:2–1971:4.

Sargent's earlier results using our alternative parameterization. This provided a welcome check of the calculations and revealed some interesting properties of the model.

Sargent assumed that we can write the unrestricted model as

$$\Delta \boldsymbol{Y}_t = \mathbf{B}\Delta \boldsymbol{Y}_{t-1} + \boldsymbol{N}_t \tag{3.12}$$

where

$$\begin{aligned}\Delta \boldsymbol{Y}_t &= (\boldsymbol{y}_t - \boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2} \ldots, \boldsymbol{y}_{t-L+1} - \boldsymbol{y}_{t-L})' \\ \boldsymbol{N}_t &= (\boldsymbol{n}_t, 0 \ldots, 0)'\end{aligned} \tag{3.13}$$

The (2×1) vector $\boldsymbol{n}_t$ is assumed to be a white noise process with contemporaneous covariance matrix $\boldsymbol{\Omega}$, and the matrix $\mathbf{B}$ has the same structure as the matrix $\mathbf{A}$ in Eq. (2.3b).

Eq. (2.4) can be first differenced, and if we take expectations at time $t-1$, we obtain

$$\begin{aligned}E_{t-1}(R_t - R_{t-1}) = \frac{1}{N} E_{t-1}\big((r_{t+N-1} - r_{t+N-2}) \\ + (r_{t+N-2} - r_{t+N-3}) + \ldots + (r_t - r_{t-1})\big)\end{aligned} \tag{3.14}$$

Once again, this can conveniently be expressed as restrictions on the first differenced bivariate process. Proceeding as before, we can rewrite Eq. (3.14) as

$$e_2'\mathbf{B}\Delta \boldsymbol{Y}_t = \frac{1}{N} e_1'(\mathbf{B} + \ldots + \mathbf{B}^N)\Delta \boldsymbol{Y}_t \tag{3.15a}$$

In general, the only guarantee that Eq. (3.15a) holds is to require that the nontrivial elements of $\mathbf{B}$ satisfy restrictions of the form

$$e_2'\mathbf{B} = \frac{1}{N} e_1'(\mathbf{B} + \ldots + \mathbf{B}^N) \tag{3.15b}$$

Because the matrix $\mathbf{B}$ has the same form as $\mathbf{A}$, all of the previous results apply, except that we can now consider having a block $\boldsymbol{\Lambda}_i$ with L (rather than $L-1$) zeros whose corresponding generalized eigenvector parameters are unrestricted.

At this point, it is very easy to get lost in the technical details and forget that it is actually a relatively simple idea which we wish to test. Fig. 1 succinctly describes what is at stake. Using the unconstrained least squares estimate of $\mathbf{B}$, we calculated the predicted changes in the long rate using both sides of Eq. (3.15a). The rather jagged line gives us the forecasted change in the long rate using the LHS of Eq. (3.15a). The very smooth curve is the average of the predicted first differences in the short rate that appears on the RHS of Eq. (3.15a). If the expectations model is correct, these two curves should differ only because of sampling error in the estimate of $\mathbf{B}$.

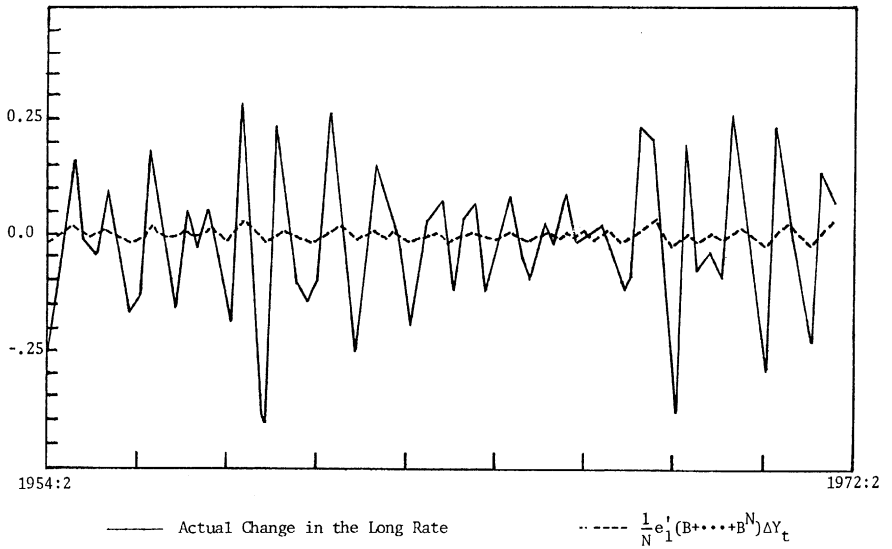

Fig. 1. Predicted and actual values of ΔR_t from the unconstrained estimate of **B**.

Of course, the two curves do appear to be quite different. Although their correlation is about 0.7, the standard deviation of the series depicted by the jagged graph is about 16 times the standard deviation of the other. This captures the popular notion that the long rate is too volatile to be consistent with the expectations model.

Unfortunately, visual inspection is not enough. While Fig. 1 gives us some idea of the kind of deviation from the expectations model that occurred in the sample, we have to appeal to a statistical theory in order to judge whether or not these deviations are "large."

The model was estimated over the period 1954:2 to 1971:4, subject to the restrictions (Eq. (3.15b)), for each of the cases $K = 0, 1 \ldots, L - 1$. Since this is a nonlinear problem and there is no guarantee of convergence to a global maximum, the choice of starting values for the parameter estimates is important. It is well known that one Newton–Raphson iteration starting from a consistent estimator is asymptotically equivalent to ML. This provides an appealing large sample argument for choosing initial estimates that are consistent, whenever possible.

In the case where the true **B** matrix is hypothesized to be nonsingular, we can choose the eigenvalues of the unrestricted least squares estimate of **B**. These are reported in Table 1, along with the corresponding eigenvectors.[29]

[29] Each entry is a pair of numbers comprised of its real and imaginary components. Note that the unconstrained estimates of the eigenvalues all fall inside the unit circle, and that they appear to be reasonably different from zero.

Table 1
Eigenvalues and eigenvectors of the least squares estimation of **B**

Eigenvalues			
$-0.6392154826D+00$	$0.1412539231D+00$	$-0.6392154826D+00$	$-0.1412539231D+00$
Eigenvectors			
$-0.2229189980D+00$	$0.1703291545D+00$	$-0.2229189980D+00$	$-0.1703291545D+00$
$0.3886437625D+00$	$-0.1805833893D+00$	$0.3886437625D+00$	$0.1805833893D+00$
$-0.6392154826D+00$	$0.1412539231D+00$	$-0.6392154826D+00$	$-0.1412539231D+00$
$0.1000000000D+01$	0.0	$0.1000000000D+01$	0.0
$-0.1142753609D+00$	$0.2102320130D+00$	$-0.1142753609D+00$	$-0.2102320130D+00$
$0.2397454004D+00$	$-0.2759117064D+00$	$0.2397454004D+00$	$0.2759117064D+00$
$-0.4485427247D+00$	$0.3325221191D+00$	$-0.4485427247D+00$	$-0.3325221191D+00$
$0.7786400787D+00$	$-0.3481394919D+00$	$0.7786400787D+00$	$0.3481394919D+00$
Eigenvalues			
$-0.4544607106D-01$	$0.7840166190D+00$	$-0.4544607106D-01$	$-0.7840166190D+00$
Eigenvectors			
$0.8371079175D-01$	$-0.4770631543D+00$	$0.8371079175D-01$	$0.4770631543D+00$
$-0.6126167135D+00$	$-0.7126094996D-01$	$-0.6126167135D+00$	$0.7126094996D-01$
$-0.4544607106D-01$	$0.7840166190D+00$	$-0.4544607106D-01$	$-0.7840166190D+00$
$0.1000000000D+01$	0.0	$0.1000000000D+01$	0.0
$0.2580392279D+00$	$-0.1162270596D+00$	$0.2580392279D+00$	$0.1162270596D+00$
$-0.1667632725D+00$	$-0.3194581420D+00$	$-0.1667632725D+00$	$0.3194581420D+00$
$-0.3938107484D+00$	$0.2355312620D+00$	$-0.3938107484D+00$	$-0.2355312620D+00$
$0.3284287434D+00$	$0.4832613789D+00$	$0.3284287434D+00$	$-0.4832613789D+00$
Eigenvalues			
$0.3359790968D+00$	$0.4684286442D+00$	$0.3359790968D+00$	$-0.4684286442D+00$
Eigenvectors			
$-0.1832410610D+00$	$0.5584628110D-01$	$-0.1832410610D+00$	$-0.5584628110D-01$
$-0.1065434413D+00$	$0.3147644656D+00$	$-0.1065434413D+00$	$-0.3147644656D+00$
$0.3359790968D+00$	$0.4684286442D+00$	$0.3359790968D+00$	$-0.4684286442D+00$
$0.1000000000D+01$	0.0	$0.1000000000D+01$	0.0
$-0.2942501505D+00$	$0.3232030928D-01$	$-0.2942501505D+00$	$-0.3232030928D-01$
$-0.2519418893D+00$	$0.4474597031D+00$	$-0.2519418893D+00$	$-0.4474597031D+00$
$0.3760245879D+00$	$0.8075473081D+00$	$0.3760245879D+00$	$-0.8075473081D+00$
$0.1518518007D+01$	$0.2864165591D+00$	$0.1518518007D+01$	$-0.2864165591D+00$
Eigenvalues			
$0.3061324594D+00$	$0.2210075909D+00$	$0.3061324594D+00$	$-0.2210075909D+00$
Eigenvectors			
$-0.1616868679D-01$	$0.5134158673D-01$	$-0.1616868679D-01$	$-0.5134158673D-01$
$0.4487272743D-01$	$0.1353151947D+00$	$0.4487272743D-01$	$-0.1353151947D+00$
$0.3061324594D+00$	$0.2210075909D+00$	$0.3061324594D+00$	$-0.2210075909D+00$
$0.1000000000D+01$	0.0	$0.1000000000D+01$	0.0
$-0.3263413672D-01$	$0.1354423376D-01$	$-0.3263413672D-01$	$-0.1354423376D-01$
$-0.4908052389D-01$	$0.7967597477D-01$	$-0.4908052389D-01$	$-0.7967597477D-01$
$0.1812449279D-01$	$0.2471816430D+00$	$0.1812449279D-01$	$-0.2471816430D+00$
$0.4221163582D+00$	$0.5026932587D+00$	$0.4221163582D+00$	$-0.5026932587D+00$

Table 2

Model	α_1	α_2	α_3	α_4	γ_1	γ_2	γ_3	γ_4	$\lvert\hat{\boldsymbol{\Sigma}}\rvert$
	β_1	β_2	β_3	β_4	δ_1	δ_2	δ_3	δ_4	
Unrestricted	−0.3663	−0.3235	0.1234	−0.0694	−0.2962	0.0203	0.2480	−0.1047	0.0163
	0.6373	0.4322	−0.3286	0.1703	0.2812	0.1200	−0.3934	0.0765	
Sargent's	−0.0717	−0.3660	−0.1465	0.0433	−0.0183	−0.0154	−0.0033	0.0014	0.0184
restricted	0.3700	0.3270	0.0995	0.0900	0.0298	0.0172	0.0063	0.0029	
Restricted									
$\mathbf{K}=0$	−0.7138	0.0077	−0.7065	0.2369	−0.1565	−0.0219	−0.1612	−0.0604	0.0895
	0.6738	−0.4035	0.7179	1.5277	−0.3247	−0.2423	0.2555	1.665	
$\mathbf{K}=1$	−0.6058	−0.0648	−0.6964	−0.0110	0.0362	0.0161	−0.1010	−0.0105	0.0958
	0.6151	−0.3619	0.8270	1.725	−0.5578	−0.3473	0.1967	1.643	
$\mathbf{K}=2$	0.0180	−0.2897	−0.1169	0.1214	0.0529	0.0377	0.0000	0.0120	0.0480
	−0.0836	−1.2949	0.1232	0.0208	−0.3587	−1.4404	−0.0141	0.0021	
$\mathbf{K}=3$	−0.4702	−0.2678	−0.2404	−0.0092	−0.3596	0.0592	−0.1515	−0.0062	0.0470
	0.6911	−1.147	0.3778	−0.0724	0.2382	−1.3220	0.3313	−0.0486	
$\mathbf{K}=4$	−0.0733	−0.3456	−0.1004	0.0975	−0.0153	−0.0123	0.0000	0.0035	0.0188
	0.3521	0.3135	0.0519	−0.0271	0.0249	0.0120	0.0008	−0.0009	

In the case where the true **B** matrix is hypothesized to have $K>0$ zero eigenvalues, we can choose the largest $(L-K)$ eigenvalues, and use the unrestricted least squares estimate of the short rate equation to solve for the implied eigenvector parameters corresponding to $\lambda=0$.

Since the least squares estimator of **B** is consistent, initial values chosen in this fashion have the advantage, by Slutsky's theorem, of also being consistent estimators, at least for the correct choice of K. Of course, in the cases where the tentatively maintained value of K is wrong, then there is no reason to believe that these chosen initial values should be close to the optimum. For example, in the hypothesized case $K=0$, it turned out that the initial values chosen gave a value of $\det(\hat{\boldsymbol{\Sigma}})$ of 1.65×10^{6}! The fact that the algorithm used in the estimation climbed smoothly (albeit slowly) to a value of $\det(\hat{\boldsymbol{\Sigma}})\simeq0.09$ seems most remarkable.

The final results are reported in Table 2. For comparison, we report the parameter estimates of the unrestricted model and Sargent's estimates of the restricted model, as well as the estimates we obtained for each of the cases $K=0,1\ldots,4$. The last column gives the value of the determinant of the covariance matrix of residuals.

Of course, the usual caveats about dealing with nonlinear models must be kept in mind. Nonetheless, several interesting results appear in Table 2.

First of all, the global ML estimates correspond to the case $K=4$.[30] The reported results imply a value of $\det(\hat{\boldsymbol{\Sigma}})$ which is slightly larger than that of

[30] Given the choice of the five local optima of the likelihood function given in the Table 2, Wald's theorem tells us to choose the one with the highest value.

Table 3

Sargent's restricted MLE	−0.0717	−0.3660	−0.1465	0.0433	0.3700	0.3270	0.0995	0.0900
of (α, β, γ, δ)	−0.0183	−0.0154	−0.0033	0.0014	0.0298	0.0172	0.0063	0.0029
Value of $(1/N)e_1'(I+$	0.317752761356$D-01$		−0.154364619860$D-01$		−0.334157786952$D-02$		0.141892663821$D-02$	
$\mathbf{B}+\ldots\mathbf{B}^{N-1})$	0.298326535924$D-01$		0.171959537979$D-01$		0.629610179904$D-02$		0.294897410895$D-02$	

$e_2'\Delta\boldsymbol{Y}_t$	$(1/N)e_1'(I+\mathbf{B}+\ldots\mathbf{B}^{N-1})\Delta\boldsymbol{Y}_t$
−0.395000000000$D+00$	−0.282589851767$D-01$
−0.350000000000$D-01$	−0.153003220712$D-01$
0.130000000000$D+00$	0.162687596329$D-01$
0.290000000000$D+00$	0.138912525039$D-01$
0.230000000000$D+00$	0.114463449006$D-01$
0.210000000000$D+00$	0.120223485500$D-01$
0.140000000000$D+00$	0.296421617040$D-01$
0.400000000000$D-01$	0.625611703906$D-02$
0.235000000000$D+00$	0.282179776302$D-02$
−0.125000000000$D+00$	−0.557159158161$D-02$
0.475000000000$D+00$	0.356039052271$D-01$
0.165000000000$D+00$	0.121757754933$D-01$
−0.120000000000$D+00$	−0.548923013734$D-02$
0.455000000000$D+00$	0.167451841934$D-01$
0.105000000000$D+00$	0.249506828886$D-01$
−0.107000000000$D+01$	−0.598482392391$D-01$
−0.360000000000$D+00$	−0.694635498431$D-01$
−0.350000000000$D-01$	0.625011424652$D-02$
0.100500000000$D+01$	0.858790267206$D-01$
0.155000000000$D+00$	−0.354640281204$D-02$
0.310000000000$D+00$	0.164061203483$D-01$
0.310000000000$D+00$	0.285956713662$D-01$
0.380000000000$D+00$	0.368746578663$D-01$
0.160000000000$D+00$	0.168043287828$D-01$
−0.840000000000$D+00$	−0.803582983148$D-01$

Sargent's reported estimates. Nonetheless, the point estimates are reassuringly close, and it appears that both our algorithm and Sargent's have converged to approximately the same solution.

Also, there appears to be a sharp difference between the $K = 4$ optimum, and the other cases. Consider once again the restrictions

$$e_2'\mathbf{B} = \frac{1}{N} e_1'(\mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^N)$$

or

$$e_2'\mathbf{P}\boldsymbol{\Lambda} = \frac{1}{N} e_1'\mathbf{P}\boldsymbol{\Lambda}(\mathbf{I} + \boldsymbol{\Lambda} + \ldots + \boldsymbol{\Lambda}^{N-1}) \tag{3.16}$$

If the number of zero roots of the characteristic polynomial is less than the lag length, we can easily verify that Eq. (3.16) is equivalent to

$$e_2'\Delta \boldsymbol{Y}_t = \frac{1}{N} e_1'(\mathbf{I} + \mathbf{B} + \ldots + \mathbf{B}^{N-1})\Delta \boldsymbol{Y}_t \tag{3.17}$$

Intuitively, Eq. (3.17) implies that we cannot make the coefficients in **B** "too small," since the RHS of Eq. (3.17) must have as much variation as the LHS. Because of the smoothing characteristics of Eq. (3.16), this will require that eigenvalues of **B** be "large," i.e., close to or outside of the unit circle. By contrast, if $K = L$, then Eq. (3.17) is not an implication of the model. The point is made in Table 3. Table 3 gives the actual changes in the long rate, and the changes "predicted" using Eq. (3.17) evaluated at the MLE of **B**, for the first 25 observations in the sample. The predicted changes usually have the right sign, but they are noticeably closer to zero and have much smaller variance than the actual changes.

4. The model in level form

In Section 3.2, it was shown that if we impose the RE restrictions on the first difference representation of the long and short rate, that they would not be rejected at conventional levels of significance. However, the model given by Eq. (2.4) was specified in terms of restrictions on the levels of $(r_t R_t)$ process. Have we lost anything by looking at Eq. (3.16) rather than at the original restrictions?

The answer turns out to be yes.

Basically, Sargent only imposed the restriction that forecast errors are uncorrelated with information available to agent's at time $t-1$. He did not require that the forecast errors be uncorrelated with the *current* levels of the short and long rate.[31] The volatility restrictions recently demonstrated by Shiller (1979) can be

[31] These particular restrictions were lost when the original restrictions (Eq. (2.4)) were first differenced and projected on I_{t-1}.

shown to be an implication of the assumption that the forecast errors are uncorrelated with the current level of the long rate.

One of the consequences of imposing the restrictions at time t is that we cannot simultaneously maintain that the RE restrictions hold and that the representation (3.13) exists, without requiring in addition that the $(r_t R_t)$ process be singular. More precisely, we claim that if the $(r_t R_t)$ process satisfies Eq. (2.4) and the matrix **B** given by Eq. (3.13) is finite dimensional, then the $(r_t R_t)$ process is singular.

The following simple example illustrates the basic idea. Define the vector $\Delta \mathbf{Y}_t = (\Delta r_t,\ \Delta R_t)'$ and $\mathbf{B} = 0$. Clearly, **B** satisfies Eq. (3.15b). Consider now the restriction on the level process given in Eq. (2.7). Since $E(\Delta \mathbf{Y}_t \mid I_t') = \mathbf{B}\Delta \mathbf{Y}_{t-1}$ by assumption, we have

$$\begin{aligned} E(r_t \mid I_t') &= r_{t-1} \\ E(R_t \mid I_t') &= R_{t-1} \end{aligned} \tag{4.1}$$

We can "undo" the first difference representation and obtain the autoregressive representation for the level process. Let $\mathbf{Y}_t = (r_t,\ R_t,\ r_{t-1},\ R_{t-1})'$. We deduce that $\boldsymbol{E}(\mathbf{Y}_t \mid I_t') = \mathbf{A}\mathbf{Y}_{t-1}$, where the matrix **A** is just

$$\mathbf{A} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{vmatrix} \tag{4.2}$$

Since the coefficients on lagged longs in the short rate equation are all zero, it is easy to see that this matrix cannot satisfy Eq. (2.7). As we have argued earlier, the interpretation to make in such a case is that the short and long rate process must be singular in order to satisfy Eq. (2.6).

This technique of "undoing" the first order representation can be used to show that the first order representation is in general inconsistent with the restrictions given by Eq. (2.7). Suppose that

$$E(\Delta r_t \mid I_{t-1}') = \sum_{i=1}^{L} a_i \Delta r_{t-i} + \sum_{i=1}^{L} b_i \Delta R_{t-1} \tag{4.3a}$$

Since $r_{t-1},\ R_{t-1} \in I_{t-1}'$, we can always rewrite this equation as

$$E(r_t \mid I_{t-1}') = \sum_{i=1}^{L+1} \alpha_i r_{t-i} + \sum_{i=1}^{L+1} \beta_i R_{t-i} \tag{4.3b}$$

where

$$\begin{aligned} &\alpha_1 = (1 + a_1) \quad \beta_1 = b_1 \\ &\alpha_i = (a_i - a_{i-1}) \quad \beta_i = (b_i - b_{i-1}) \qquad i = 2, \ldots L \\ &\alpha_{L+1} = -a_L \quad \beta_{L+1} = -b_L \end{aligned}$$

Of course, we can do exactly the same thing with the long rate equation so that we can construct the level representation directly from the matrix **B**. Therefore, there are certain restrictions implied on the resulting matrix **A**. The fact that it is inconsistent with Eq. (2.7) is an important example of the previous statement that it may not be possible to satisfy Eq. (2.7) as well as other a priori restrictions.

The matrix **B** of the first difference representation and the implied **A** matrix of the level process are intimately related. It is easy to show that if λ is an eigenvalue of **B**, then it is also an eigenvalue of **A**. Moreover, the corresponding eigenvector parameters are also the same. The main difference is that **A** has two unit roots in addition to the roots of **B**. To check if the implied **A** matrix satisfies Eq. (2.7), therefore, we only have to consider the generalized eigenvectors corresponding to these unit roots to see if they satisfy the restrictions analogous to Eq. (3.6).

Doing so produces the following contradiction. The matrix **A**, which we have constructed, can easily be shown to have two independent eigenvectors corresponding to the unit root. However, from Section 3, we know that unless $\lambda = 0$, there cannot be more than one independent eigenvector corresponding to each latent root, and moreover, we know exactly what the eigenvector must be. We conclude that the matrix **A** which we have constructed from the first difference representation cannot solve Eq. (2.7).

In terms of the interpretation given in Section 3.1, the fact that **A** has two independent eigenvectors corresponding to the unit root implies the existence of a component Z_1 which contributes only to the short rate process, and a component Z_2 which contributes only to the long rate. However, the long rate is hypothesized to be the average of optimally forecasted future short rates. Therefore, Z_2 must be equivalent to the optimal forecast of $S(\lambda_i)Z_1$. The only way this can happen is if the component Z_2 is proportional to Z_1 at all points in time. Since $\mathbf{P}^{-1}$ is nonsingular, we are forced to conclude that $\boldsymbol{Y}_t$ must be a singular process.

It is possible to demonstrate the inconsistency in a somewhat more direct manner.[32] If the matrix **A** has two independent eigenvectors corresponding to the unit root, they can always be written as[33]

$$J = \begin{vmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 0 & 1 \end{vmatrix}$$

[32] I am indebted to Bob Shiller for pointing out this line of reasoning to me.

[33] Recall that the eigenvectors must be of the form $\tilde{\boldsymbol{\lambda}}_i \otimes \boldsymbol{p}_i$. Since $\lambda_i = 1$, $\tilde{\lambda}_i = (1, 1\ldots, 1)'$. To construct the eigenvectors corresponding in the unit root, we get to choose any $\boldsymbol{p}_i$, $\boldsymbol{p}_{i+1}$ vectors that allow us to span a two-dimensional space, so in particular we can pick $\boldsymbol{p}_i = (1,0)$, and $\boldsymbol{p}_{i+1} = (0,1)$.

Postmultiplying both sides of Eq. (2.7) gives us

$$e_2'\mathbf{J} = \frac{1}{N}e_1'(\mathbf{I} + \mathbf{A} + \ldots \mathbf{A}^{N-1})\mathbf{J} = \frac{1}{N}e_1'(\mathbf{J} + \mathbf{J} + \ldots + \mathbf{J}) = e_1'\mathbf{J}$$

or

$$(0,1) = (1,0)$$

Intuitively, it is easy to see that there is a problem with the first difference representation. The assumption that the representation (3.13) exists gives us four restrictions on the constructed matrix **A**, namely $\sum\alpha_i = \sum\delta_i = 1$ and $\sum\beta_i = \sum\gamma_i = 0$. This is enough to determine the two extra eigenvalues and eigenvector parameters of **A** uniquely, given only knowledge of **B**. However, Eq. (2.7) gives us two extra restrictions. With six restrictions and only four more parameters to determine, it is not surprising no solution exists.

The main consequence of this result is that we should consider the (finite) first difference representation basically incompatible with the expectations model defined by Eq. (2.4). If it were the true representation, we would be able to find a linear combination of $\boldsymbol{Y}_t$ that equals zero with probability one in all time periods.

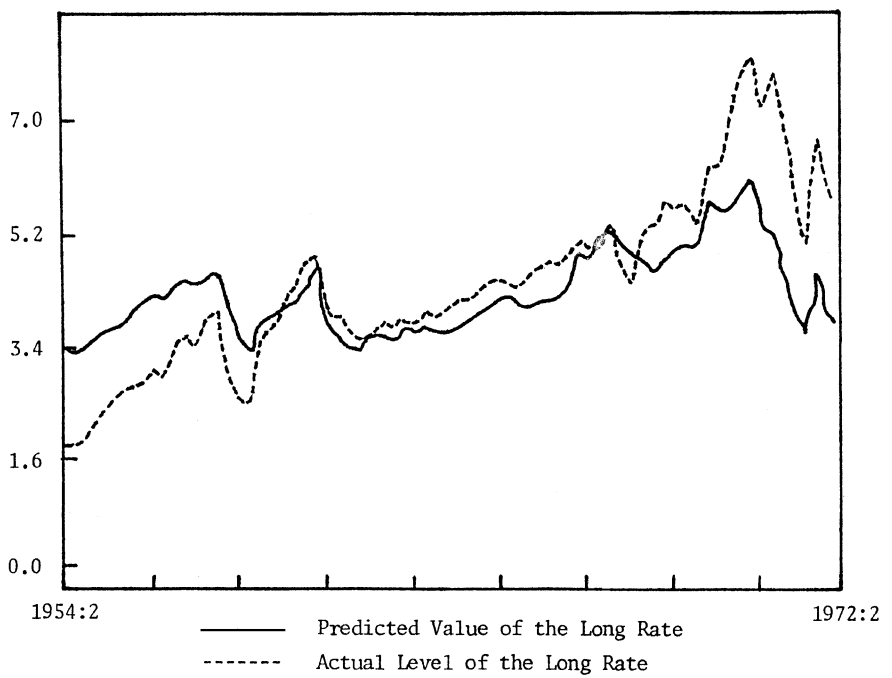

Fig. 2. Predicted and actual values of R_t from the unconstrained estimate of **A**.

Table 4

Model	α_1	α_2	α_3	α_4	α_5	γ_1	γ_2	γ_3	γ_4	γ_5	$\lvert\hat{\Sigma}\rvert$
	β_1	β_2	β_3	β_4	β_5	δ_1	δ_2	δ_3	δ_4	δ_5	
Unrestricted	0.5436	0.0002	0.43253	−0.15298	0.0956	−0.2509	0.2992	0.2956	−0.3295	0.1753	0.012
	0.69481	−0.1755	−0.7131	0.4739	−0.3069	1.189	−0.1838	−0.5673	0.4552	−0.1547	
Restricted	1.006	−0.253	0.149	0.051	−0.046	−0.105	0.005	−0.008	−0.0002	0.003	
2 unit roots,											0.020
4 zero roots	0.261	−0.1405	−0.1965	0.2673	−0.099	1.09	0.009	0.001	−0.009	0.005	
Sargent's	0.9283	−0.2943	0.2195	0.1898	−0.0433	−0.183	0.003	0.012	0.0047	−0.0014	
restricted											0.018
"solved out"	0.370	−0.043	−0.2275	−0.0950	−0.0900	1.03	−0.0126	−0.011	−0.003	−0.003	

If we are unable to find such a linear combination, we can reject the combined maintained hypothesis that the first differences representation exists and that the rational expectation restrictions hold with probability one.[34]

Using the alternative parameterization discussed in Section 3, there is nothing preventing us from estimating the model in level form and imposing all of the RE restrictions. Once again, it is useful to look at a graph of what's at stake. Fig. 2 plots the actual yield to maturity on our long bonds, and the predicted average of future short rates. Again, the expectations model claims that these two curves differ only because of sampling error in the least squares estimate of the matrix **A**. In fact, the two curves are remarkably similar over the middle range of the sample period, and do exhibit similar movements throughout.[35]

The various estimates are presented in Table 4. The first row gives the least squares estimates of the model in level form. The second row gives the model estimated subject to Eq. (2.7) under the maintained hypothesis that two of the roots of the autoregression are unity, and four of them are zero. We chose to set the two roots to unity in order to make the model roughly comparable to Sargent's.[36] The decision to fix four of the eigenvalues to zero stemmed from the experience gained in estimating the model in first differences. Of course, relaxing these restrictions can only serve to improve the fit of the model, and the estimates presented would not be rejected at the usual levels of significance.

For comparison, the solved out coefficients from Sargent's model are also presented in the bottom row.

Several comments about Table 4 seem in order. First of all, the similarity between the estimates obtained and Sargent's earlier estimates is striking, even though the latter do not satisfy all of the RE restrictions. For all practical purposes, imposing the extra restrictions has not changed the estimates. Also striking is the similarity between these results and what we might expect if $(\Delta r_t, \Delta R_t)$ was really just white noise.

5. The Wald test

The discussion so far has centered upon methods of imposing exactly the restrictions implied by the RE hypothesis. There are basically two closely related motivations for imposing these restrictions.

[34] Note the important role which the finiteness of the dimension of **A** plays in this argument. To be precise, there really are no testable implications as long as L, the order of the vector autoregression, is assumed to be larger than the sample size.

[35] It should be kept in mind that the discrepancies noted in the first differences of these graphs are implicitly still there.

[36] Of course, there is only one independent eigenvector corresponding to these unit roots. It is constrained by Eq. (3.6).

The first is to obtain more efficient estimates of the matrix **A** in the vector autoregression. While we could imagine situations where these parameters are of direct interest, in this paper, the coefficients in **A** do not represent any structural parameters of economic importance; therefore, it is difficult to justify expending much effort in their estimation.

The main reason for trying to find simple ways to impose the RE restrictions is to facilitate a likelihood ratio test of them. Likelihood ratio tests are by far the most common tests used by economists. Under appropriate assumptions, they are known to have desirable large sample properties. However, there are alternatives.

Two other members of what some econometricians refer to as "the holy trinity" are the Lagrange multiplier (or Efficient Score) and Wald tests.[37] These two other tests have in large samples the same distribution under the null hypothesis and the same power against local alternatives as the likelihood ratio test. According to the standard large sample criteria used by econometricians, therefore, there is no reason for preferring any one of these three tests on statistical grounds. However, the computational burden associated with each of these tests can be quite different.

The likelihood ratio test requires us to maximize the likelihood function over both the restricted and unrestricted parameter spaces. By contrast, the Lagrange multiplier test only requires that we maximize the likelihood function under the null, and the Wald test only that we maximize it under the alternate hypothesis. Since the unrestricted model reduces to multivariate regression, the unrestricted likelihood function is maximized at the least squares estimates of $(\alpha, \beta, \gamma, \delta)$. This makes consideration of the Wald test very attractive. The basic features of this test are well described elsewhere, but some of the more relevant results will be repeated here for convenience.

First, we introduce some new notation. Define the vector $\boldsymbol{\theta}' = (\alpha, \beta, \gamma, \delta)$. The restrictions given by Eq. (2.7) can be rewritten as $\boldsymbol{g}(\boldsymbol{\theta}) = 0$, where $\boldsymbol{g}(\boldsymbol{\theta})$ is a $(2L \times 1)$ vector given by

$$\boldsymbol{g}(\boldsymbol{\theta}) = e_1 - (I + \mathbf{A} + \ldots + \mathbf{A}^{N-1})' e_2 \tag{5.1}$$

The covariance matrix of the asymptotic distribution of the least squares estimates, $\boldsymbol{\theta}_{\mathrm{LS}}$, is given by $\mathbf{H}^{-1} = 1/T(\boldsymbol{\Sigma} \otimes M_{YY}^{-1})$ where $M_{YY} = \operatorname{plim} (1/T)\boldsymbol{\Sigma} Y_{t-1} Y_{t-1}'$. Finally, define $R(\boldsymbol{\theta}_0) = ((\mathrm{d}\boldsymbol{g}(\boldsymbol{\theta}))/(\mathrm{d}\boldsymbol{\theta}))(\boldsymbol{\theta} = \boldsymbol{\theta}_0)$, i.e., the $(4L \times 2L)$ matrix of partial derivatives whose (i,j)th element is $\mathrm{d}\boldsymbol{g}_j(\boldsymbol{\theta})/\mathrm{d}\boldsymbol{\theta}_i$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Perhaps the most illuminating interpretation of the properties of the Wald test stresses its similarity to the likelihood ratio test of the linear restrictions $R\boldsymbol{\theta} = r$. Assume for the time being that $\mathbf{H}^{-1}$, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$, is known. The maximum value of the likelihood function over the restricted parameter space is found by minimizing the quadratic form $\mathrm{LR} = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\mathrm{LS}})' \mathbf{H} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\mathrm{LS}})$ subject to $\boldsymbol{g}(\boldsymbol{\theta}) = 0$.

[37] For a discussion of these large sample tests and their relationship to each other, Cox and Hinkley (1974), Chap. 9.

The problem is depicted graphically in Fig. 3a,b. The level contours of LR generate ellipsoids which correspond to the standard family of confidence intervals for $\boldsymbol{\theta}$, based on the chi-squared distribution with $4L$ degrees of freedom. The level contours of $\boldsymbol{g}(\boldsymbol{\theta})$ can be very simple as in Fig. 3a, or more complicated as in Fig. 3b. Nonetheless, what has to be done is to search along the contour $\boldsymbol{g}(\boldsymbol{\theta}) = 0$ to find the value, call it $\hat{\boldsymbol{\theta}}_{\mathrm{R}}$, on the ellipsoid closest to the unrestricted estimate. The minimum value which LR obtains is just the usual likelihood ratio test statistic, so that the latter can be thought of as a measure of the distance between $\hat{\boldsymbol{\theta}}_{\mathrm{R}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$, distance being measured in the metric of **H**. The likelihood ratio test, therefore, amounts to finding, $\hat{\boldsymbol{\theta}}_{\mathrm{R}}$ and checking to see if it lies inside the ellipsoid of the chosen $(1-\alpha)$ percent confidence interval.

Complications arise because it may be difficult to search along the contours of $\boldsymbol{g}(\boldsymbol{\theta}) = 0$. The Wald approach gets around this problem by replacing the contours $\boldsymbol{g}(\boldsymbol{\theta}) = 0$ by the linearized approximation

$$\tilde{\boldsymbol{g}}(\boldsymbol{\theta}) = \boldsymbol{g}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right) + \mathbf{R}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right) \tag{5.2}$$

We can then use closed form expressions to find the point $\hat{\boldsymbol{\theta}}_{\mathrm{W}}$ which minimizes LR subject to $\tilde{\boldsymbol{g}}(\boldsymbol{\theta}) = 0$, namely

$$\hat{\boldsymbol{\theta}}_{\mathrm{W}} = \hat{\boldsymbol{\theta}}_{\mathrm{LS}} - \mathbf{H}^{-1}\mathbf{R}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right)'\left(\mathbf{R}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right)\mathbf{H}^{-1}\mathbf{R}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right)'\right)^{-1}\boldsymbol{g}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right) \tag{5.3}$$

and the value of the quadratic form reduces to

$$\mathrm{LR}\left(\hat{\boldsymbol{\theta}}_{\mathrm{W}}\right) = \boldsymbol{g}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right)'\left(\mathbf{R}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right)\mathbf{H}^{-1}\mathbf{R}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right)\right)^{-1}\boldsymbol{g}\left(\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\right) \tag{5.3}$$

Under the null hypothesis, both $\mathrm{LR}(\hat{\boldsymbol{\theta}}_{\mathrm{R}})$ and $\mathrm{LR}(\hat{\boldsymbol{\theta}}_{\mathrm{W}})$ are distributed asymptotically as chi-square with degrees of freedom equal to the rank of $\mathbf{R}(\boldsymbol{\theta})$. Further-

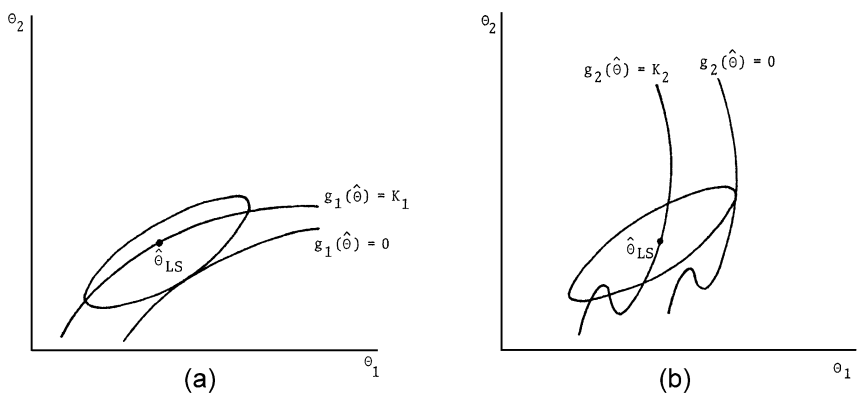

Fig. 3.

Table 5
Wald estimates of the restricted model

$\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$	−0.3575	−0.3393	0.1335	−0.0708	0.6243	0.4450	−0.3333	0.1709
	−0.2832	0.0092	0.2548	−0.0999	0.2649	0.1274	−0.3972	−0.0698
$\boldsymbol{g}(\hat{\boldsymbol{\theta}}_{\mathrm{LS}})$	0.2619	−0.1266	−0.2488	0.0951	−0.2372	−0.1248	0.3835	−0.0629
$\hat{\boldsymbol{\theta}}_{\mathrm{W}}$	−0.0767	−0.3677	−0.1430	0.0357	0.3749	0.3300	0.0987	0.1003
	−0.0188	−0.0141	−0.0026	0.0006	0.0298	0.0159	0.0048	0.0035
$\boldsymbol{g}(\hat{\boldsymbol{\theta}}_{\mathrm{W}})$	0.0003	−0.0014	−0.0009	0.0005	0.0003	0.0015	0.0018	−0.0003

more, the linearized maximum likelihood estimator (LMLE), $\hat{\boldsymbol{\theta}}_{\mathrm{W}}$, has the same asymptotic distribution as $\hat{\boldsymbol{\theta}}_{\mathrm{R}}$. For a given data set, however, the two estimators can be quite different. Their similarity will depend, of course, upon the adequacy of the linear approximation $\tilde{\boldsymbol{g}}(\boldsymbol{\theta})$.

Some results are collected in Table 5 for the first differenced model discussed in Section 3.2. The first two rows give the computed estimates of $(\alpha,\ \beta,\ \gamma,\ \delta)$, and the value of $\boldsymbol{g}(\hat{\boldsymbol{\theta}}_{\mathrm{LS}})$. The next two rows give the value of the linearized maximum likelihood estimator and the value of $\boldsymbol{g}(\hat{\boldsymbol{\theta}}_{\mathrm{W}})$.

The Wald statistic works out to be about 8.89, which is very close to the likelihood ratio statistic reported in Table 2. Moreover, the actual point estimates are extremely close to those found by maximizing the likelihood function subject to the exact RE restrictions.

6. Some extensions

This section contains a collection of topics which indicate directions for future research.

6.1. A larger information set

Conceptually, it is relatively straightforward to add other variables to the information set and evaluate Eq. (2.5). Suppose we wish to include another variable, call it x_t, besides the short and long rate in the information set I_t'. If we reinterpret the vector $\boldsymbol{y}_t$ as $(r_t,\ R_t,\ x_t)'$, the results discussed earlier generalize immediately. The eigenvectors of the now $(3L \times 3L)$ matrix $\mathbf{A}$ have the familiar structure $\bar{\boldsymbol{\lambda}} \otimes \boldsymbol{p}_i$, where $\boldsymbol{p}_i = (p_{1i},\ p_{2i},\ p_{3i})'$. Furthermore, the restrictions remain

$$\lambda_i^{L-1} p_{2i} = \lambda_i^{L-1} p_{1i} \sum_{i=0}^{N-1} \lambda_i^k \qquad \lambda_i \neq 0 \tag{6.1}$$

An interesting special case arises when x_t represents the yield to maturity on some other bond, say of maturity N'. Then, alongside the restrictions (6.1), we can also impose the restrictions

$$\lambda_i^{L-1} p_{3i} = \lambda_i^{L-1} p_{1i} \sum_{i=0}^{N'-1} \lambda_i^k \qquad \lambda_i \neq 0 \tag{6.2}$$

It must be confessed, however, that while conceptually, the extension of the information set to include variables other than the short and long rate is straightforward, the computational problems involved in dealing with zero eigenvalues and the many possible structures of the matrix **A** make the estimation algorithm used in this paper extremely difficult to follow.

6.2. Testing against specific alternatives

A test of restrictions (2.7) can be thought of as a test of the hypothesis that the ex post forecast error is uncorrelated with current and lagged values of the short and long rates. The likelihood ratio and Wald tests described earlier do not emphasize any particular kind of deviation from the null hypothesis. That is to say, they are not designed to treat any particular kind of departure from the null as more informative than others. Good statistical methodology, however, would have us describe the alternatives which we consider most plausible and then construct tests which are as sensitive as possible in detecting differences between the maintained hypothesis and these particular alternatives.

It is admittedly difficult to think about what these alternatives should be. The sorts of models that come to mind are stories about the behaviour of risk averse agents facing uncertainty and transactions costs. One cannot easily translate these models into statements about the coefficients of a vector autoregression.

Nonetheless, there may be good reasons to believe that the current levels of the long and short rate, rather than lagged values, should be most useful in predicting the ex post forecast error. Shiller (1979) has shown that the excessive volatility associated with the holding period yields of long bonds is a consequence of a high correlation between the ex post forecast error and the current level of the long rate.

A nice feature of emphasizing these particular deviations is that it makes the choice of the lag length in the autoregression less important. Fixing L at a large number gives some robustness to the test procedures described earlier in the paper. However, if we make L too large, we will undoubtedly introduce small eigenvalues into the least squares estimate of the matrix **A**. The rational expectations restrictions (2.7) corresponding to these eigenvalues will be satisfied rather easily in practice by setting them equal to zero. A straightforward likelihood ratio test would not reject the null hypothesis unless the deviations in the other directions were correspondingly larger.

In terms of the matrix **A**, the particular subset of restrictions we propose to test can be written as

$$(0,1) = \frac{1}{N} e_1'(\mathbf{I} + \mathbf{A} + \ldots + \mathbf{A}^{N-1})(e_1 e_2) \tag{6.3}$$

It is inconvenient to test Eq. (6.3) using the approach suggested in Section 3. However, it is very easy to test Eq. (6.3) using the Wald statistic. For the first differenced model described in Section (3.2), the marginal significance level of the test of all the restrictions in Eq. (3.15b) is approximately equal to 0.4. If we test only the first two restrictions analogous to Eq. (6.3), the calculated Wald statistic works out to be about 4.85. The marginal significance level of this last test statistic is therefore just under 0.1. Therefore, we would still not reject the restrictions at conventional levels of significance, but it appears that concentrating our attention on these particular alternatives does make the model somewhat more suspect.

7. Conclusion

This paper demonstrates the feasibility of dealing directly with the Jordan canonical form of a vector autoregression.

One of the main advantages of using this alternative parameterization is the possibility of obtaining closed form expressions for the sorts of restrictions which arise in currently popular rational expectations models. In our particular application, the closed form expressions make it possible to understand better the nature of the restrictions implied by the expectations model of the term structure. In particular, we learned that the nature of the restrictions depends rather crucially on whether or not some of the roots of the characteristic polynomial of the vector autoregression are zero. Also, it was possible to demonstrate that the bivariate finite first difference representation suggested by Sargent has certain technical drawbacks, which make it inappropriate to use in conjunction with the expectations model.

From a computational point of view, it is cumbersome to work directly with the Jordan decomposition. Some of the associated expense can be reduced by avoiding complex arithmetic, exploiting the special structure of the matrix **P** in calculating its inverse, etc. However, the main expense comes from the need to deal carefully with the many possible structures of the matrix **A** in the vector autoregression. With prior knowledge of the structure of **A**, or of the permissible values of the characteristic roots, it may be possible to do much better. Techniques for dealing with the general case remain a topic for future research.

One of the more encouraging results in this paper is the splendid performance of the Wald test and estimator. Given the flexibility and computational advantage of these procedures, they deserve to be used more often by researchers, at least in the preliminary analysis of a model.

The main conclusion of this paper is that imposing the restriction that errors in forecasting the holding period yield of long and short bonds be uncorrelated with current values of the long and short rates does not materially affect the conclusions reached in Sargent's earlier study. For the particular data set examined, at least the expectations model of the term structure fits the facts quite well.

Acknowledgements

My thanks are given to Geert Bekaert for inviting me to submit my paper after all these years, and to Ben Friedman and Olivier Blanchard for their encouragement and support.

References

Bellman, R., 1970. Introduction to Matrix Analysis, 2nd edn. McGraw-Hill, New York.

Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman & Hall, London.

Gantmacher, F.R., 1959. The Theory of Matrices, vol. 1. Chelsea Publishing, New York.

Lucas, R.E., Jr., 1976. Econometric policy evaluation: a critique. In: Brunner, K., Meltzer, A. (Eds.), The Phillips Curve and Labor Markets, vol. 1 of the Carnegie-Rochester Conferences on Public Policy, a supplementary series to the Journal of Monetary Economics. North-Holland, Amsterdam, pp. 19–46.

Melino, A., 1983. Essays on Estimation and Interference in Linear Rational Expectations Models. PhD dissertation, Harvard University.

Sargent, T.J., 1979. A note on maximum likelihood estimation of the rational expectations model of the term structure. Journal of Monetary Economics 5 (1), 133–143 (January).

Shiller, R.J., 1979. The volatility of long-term interest rates and expectations models of the term structure. Journal of Political Economy 87 (6), 1190–1219 (December).

Singleton, K., 1980. Expectations models of the term structure and implied variance bounds. Journal of Political Economy 88 (6), 1159–1176 (December).

Wald, A., 1943. Tests of statistical hypothesis concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 54, 426–482.