Testable implications of affine term structure models

James D. Hamilton\(^a\)*, Jing Cynthia Wu\(^b\)

\(^a\) Department of Economics, University of California, San Diego, United States
\(^b\) Booth School of Business, University of Chicago, United States

**Abstract**

Affine term structure models have been used to address a wide range of questions in macroeconomics and finance. This paper investigates a number of their testable implications which have not previously been explored. We show that the assumption that certain specified yields are priced without error is testable, and find that the implied measurement or specification error exhibits serial correlation in all of the possible formulations investigated here. We further find that the predictions of these models for the average levels of different interest rates are inconsistent with the observed data, and propose a more general specification that is not rejected by the data.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Affine term structure models have become a fundamental tool for empirical research in macroeconomics and finance on the term structure of interest rates. The appeal of the framework comes from the closed-form solutions it provides for bond and bond option prices under the assumption that there are no possibilities for risk-free arbitrage (Duffie et al., 2000). ATSM have been used for purposes such as measuring risk premia (Duffee, 2002; Cochrane and Piazzesi, 2009), studying the effect of macroeconomic developments on the term structure (Ang and Piazzesi, 2003; Beechey and Wright, 2009; Bauer, 2011), the role of monetary policy (Rudebusch and Wu, 2008), explaining the bond-yield “conundrum” of 2004–2005 (Rudebusch et al., 2006), inferring market expectations of inflation (Christensen et al., 2010), and evaluating the effects of the extraordinary central bank interventions during the financial crisis (Christensen et al., 2009; Smith, 2010; Hamilton and Wu, 2012a). Gürgaynak and Wright (2012) and Rudebusch (2010) provide useful surveys of this literature.

Clive Granger’s primary interest was not in a model’s theoretical elegance, but instead in its practical relevance. He would always want to know whether the framework generates useful forecasts, and whether the properties of those forecasts could be used to test some of the model’s implicit assumptions. To be sure, forecasting interest rates has been one important goal for many users of ATSM. Improved forecasts are cited by Ang and Piazzesi (2003) as an important reason for including observed macroeconomic factors in the model, and by Christensen et al. (2011) as an advantage of their dynamic Nelson–Siegel specification.\(^1\) And comparing the fit of a broad class of different models has been attempted by Dai and Singleton (2000), Hong and Li (2005) and Pericoli and Taboga (2008). However, as implemented by these researchers, making these comparisons is an arduous process requiring numerical estimation of highly nonlinear models on ill-behaved likelihood surfaces. As a result, previous researchers have overlooked some of the basic empirical implications of these models that are quite easy to test empirically.

In a companion paper (Hamilton and Wu, 2012b), we note that an important subset of ATSMs imply a restricted vector autoregression in observable variables. These restrictions take two forms: (1) nonlinear restrictions on the VAR coefficients implied by the model, and (2) blocks of zero coefficients. In this paper, we test the first class of restrictions using the \(\chi^2\) test developed by Hamilton and Wu (2012b), and note that the second class of restrictions often take the form of simple and easily testable Granger-causality restrictions, and indeed provide an excellent illustration of Granger’s (1969) proposal that testing such forecasting implications can often be a very useful tool for evaluating a model.

We apply these tests to the data and find that the assumptions that are routinely invoked in these models can in fact be routinely rejected. We show that the assumption that certain specified yields are priced without error is testable, and find that the implied measurement or specification error exhibits serial correlation in

\(^{1}\text{We thank Jonathan Wright and anonymous referees for helpful comments on an earlier draft of this paper.}\)

\(*\text{Corresponding author.}\

E-mail addresses: jhamilton@ucsd.edu (J.D. Hamilton), Cynthia.Wu@chicagobooth.edu (J.C. Wu).

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.\

http://dx.doi.org/10.1016/j.jeconom.2013.08.024
of the yields on n others should have an $R^2$ of unity. Although such an $R^2$ is not actually unity, it can be quite high, and this observation motivates the claim that a small number m of factors might be used to give an excellent prediction of any bond yield. One common approach is to suppose that there are m linear combinations of $Y_t$ for which (6) holds exactly.

$$Y_{1t} = A_1 + B_1 F_t$$

(10)

where the $(m \times 1)$ vector $Y_{1t}$ is given by $Y_{1t} = H_1 Y_t$ for $H_1$ an $(m \times N)$ matrix, $A_1 = H_1 A$, and $B_1 = H_1 B$. The matrix $H_1$ might simply select a subset of m particular yields (e.g., Chen and Scott, 1993; Ang and Piazzesi, 2003), or alternatively could be interpreted as the matrix that defines the first m principal components of $Y_t$ (e.g., Joslin et al., 2011). The remaining $N_e = N - m$ yields are assumed to be priced with error,

$$Y_{2t} = A_2 + B_2 F_t + u_{2t}$$

(11)

for $u_{2t}$ an $(N_e \times 1)$ vector of measurement or specification errors, $Y_{2t} = H_2 Y_t$, $A_2 = H_2 A$, and $B_2 = H_2 B$ for $H_2$ $(N_e \times N)$. The measurement errors have invariably been regarded as serially and mutually independent, $u_{2t} \sim$ i.i.d. $N(0, \Sigma_u, \Sigma_u')$ for $\Sigma_u$ a diagonal matrix, and with the sequence $\{u_{2t}\}$ assumed to be independent of the factor innovations $\{u_t\}$ in (1).

3. Testable implications when only yield data are used

In this section we consider the popular class of models in which the entire vector of factors $F_t$ is treated as observed only through the yields themselves. We first describe the implications for the underlying VAR in $Y_t$, and then investigate tests of the various restrictions.

3.1. VAR representation

As in Hamilton and Wu (2012b), we premultiply (1) by $B_1$.

$$B_1 F_{t+1} = B_1 c + B_1 \rho B_1^{-1} B_1 F_t + B_1 \Sigma u_{t+1}.$$  

Adding $A_1$ to both sides and using (10),

$$Y_{1,t+1} = A_1^* + \phi_1^* Y_{1t} + u_{1,t+1}.$$  

(12)

$$A_1^* = A_1 + B_1 c - B_1 \rho B_1^{-1} A_1.$$  

(13)

$$\phi_1^* = B_1 \rho B_1^{-1}.$$  

(14)

$$u_{1,t+1} = B_1 \Sigma u_{t+1}.$$  

(15)

Similar operations on (11) produce

$$Y_{2t} = A_2^* + \phi_2^* Y_{1t} + u_{2t}.$$  

(16)

$$A_2^* = A_2 - B_2 \rho B_2^{-1} A_1.$$  

(17)

$$\phi_2^* = B_2 \rho B_2^{-1}.$$  

(18)

for $u_{2t}$ the identical error as in (11).

Under the assumptions made above for $u_t$ and $u_{2t}$, the error $u_{1,t+1}$ in (12) is uncorrelated with $\{Y_t, Y_{1,t-1}, \ldots\}$, and $u_{2t}$ in (16) is uncorrelated with $\{Y_{1,t-1}, Y_{1,t-2}, \ldots\}$. Hence although the nonlinear recursions that define the ATSM are quite complicated, the fundamental structure is very simple—the ATSM is simply a vector autoregression for $(Y_{1t}', Y_{1t}')$ that is subject to a variety of restrictions. A number of these restrictions are quite simple to test without using the core equations (2) and (4), as we now discuss.

2 Duffee (2011a) has also noted the substantial serial correlation of measurement errors.
3.2 Granger-causality tests: Y_t

Eqs. (12) and (16) are a special case of a VAR(1), whose first block in the absence of restrictions would take the form

$$Y_{1t} = A_1^* + \phi_{11} Y_{1,t-1} + \phi_{12} Y_{2,t-1} + u_{1t}. \quad (19)$$

In other words, the ATSM implies that the yields priced with error Y_2 should not Granger-cause the yields priced without error Y_1. Since the coefficients of this unrestricted VAR can be estimated using OLS equation by equation, this is an extremely straightforward hypothesis to test.

We test this implication using end-of-month constant-maturity Treasury yields taken from the daily FRED database of the Federal Reserve Bank of St. Louis, using maturities of 3 months, 6 months, 1 year, 2 years, 3 years, 5 years, 7 years and 10 years. All the in-sample estimation was based on the subsample from 1983:M1 to 2002:M7, with the subsequent 60 months (2002:M8–2007:M7) reserved for out-of-sample exercises.³

For our baseline example, we use m = 3 factors and suppose that 3 yields – namely the 6-month, 2-year, and 10-year yields – are priced without error (Y_{1t} = (y_{6m}, y_{2y}, y_{10y})'), while the other yields are priced with error (Y_{2t} = (y_{6m}, y_{2y}, y_{7y}, y_{5y}, y_{3y}, y_{1y})'). The first row of Table 1 reports tests for Granger-causality from Y_2 to Y_1 for this specification. An F-test of the null hypothesis that the first row of \phi_{12} is zero (in other words, that Y_{2,t-1} does not help predict the 6-month yield) leads to strong rejection with a p-value of 0.006. Analogous tests that the second and third rows of \phi_{12} are zero (Y_{2,t-1} does not predict Y_{9,t} or Y_{10,t}) fail to reject with p-values of 0.198 and 0.204. A likelihood ratio test with Sims’ small-sample correction⁴ of the null hypothesis that all 15 elements of \phi_{12} are zero leads to very clear rejection (last column of row 1).

This test makes apparent that the specification of which yields are assumed to be priced without error is not an arbitrary normalization, but instead is a testable restriction. If Y_{6m} is priced with error, it should contain no information about the factors beyond that contained in Y_{10y}, and therefore should not help to predict Y_{10y}. If some maturities are more helpful than others for forecasting, those are the ones we would want to include in Y_1 for the ATSM to be consistent with the data. In the subsequent rows of Table 1 we report analogous F-tests and likelihood ratio tests for each of the \binom{8}{3} = 56 possible choices we could have made for the 3 yields to include in Y_1. It turns out that every single possible specification of Y_1t is inconsistent with the data according to the likelihood ratio test.

Granger (1980) expressed the view that one wants with these tests to consider true predictive power, which may be different from the ability to fit a given observed sample of data. For this reason, Granger stressed the importance of out-of-sample evaluation. In this spirit, we estimated (19) for t = 1, 2, . . . , T and used the resulting coefficients and values of Y_{1t} and Y_{2t} to predict the value of Y_{1,t+1}, whose ith element we denote \hat{y}_{1,t+1} and associated forecast error \hat{\epsilon}_{i,t+1}. We also estimated the restricted regressions with \phi_{12} = 0 to calculate a restricted forecast \hat{y}_{1,t+1} and error \hat{\epsilon}_{i,t+1}. We then increased the sample size by one to generate \hat{y}_{1,t+2} and \hat{\epsilon}_{i,t+2}, and repeated this process for T + 1, T + 2, . . . , T + R. The columns in Table 2 report the percent improvement in post-sample mean squared error.

<table>
<thead>
<tr>
<th>Specification of Y_{1t}</th>
<th>Granger-causality tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
</tr>
<tr>
<td>6m, 2y, 10y</td>
<td>0.006</td>
</tr>
<tr>
<td>6m, 3y, 10y</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 2y, 7y</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 2y, 5y</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 2y, 3y</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 2y, 1y</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 10y, 3m</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 10y, 5y</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 10y, 7y</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 10y, 9y</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 10y, 11y</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 10y, 13y</td>
<td>0.005</td>
</tr>
<tr>
<td>6m, 10y, 15y</td>
<td>0.005</td>
</tr>
</tbody>
</table>

³ We have also repeated many of the calculations reported below using the alternative measures of interest rates developed by Gürkaynak et al. (2007) and came up with broadly similar results.

⁴ Let \hat{u}_t denote the vector of OLS residuals from estimation of (19) over t = 1, . . . , T and \hat{\epsilon}_t = T^{-1} \sum_{t=1}^{T} \hat{u}_t. Let \tilde{u}_t denote the vector of OLS residuals when Y_{2,t-1} is dropped from the equation with \tilde{\hat{\epsilon}}_t = T^{-1} \sum_{t=1}^{T} \tilde{u}_t. Then as in Hamilton (1994), equation [11.3.14], (\bar{T} = N - 1)(\log \hat{\delta}_t - \log \hat{\delta}_1) is approximately \chi^2(m = N - m) for N the dimension of Y_1 and m the dimension of Y_2. All system-wide likelihood ratio tests reported in this paper use this small-sample correction, with the exception of Table 9, in which there are differing degrees of freedom across equations.
for $i$ corresponding to the first, second, or third element of $Y_{1t}$ for each of the 56 possible choices of $Y_{1t}$. For example, inclusion of $Y_{2,1,t}$ leads to a 25% out-of-sample improvement in forecasting the 6-month yield and an 8% improvement for the 2-year yield for $Y_{1t} = (Y_{2t}, Y_{24t}, Y_{120t})$.

Clark and West (2007) discussed the statistical significance of such post-sample comparisons, noting that even if the null hypothesis is false (that is, even if $Y_{2,1,t}$ actually is helpful in predicting $Y_{1t}$) we might expect the above statistic to be negative as a result of sampling uncertainty. They proposed a test statistic that corrects for this which, while not asymptotically normal, seems to be reasonably well approximated by the $N(0, 1)$ distribution,

$$C = \frac{\sqrt{R}}{R - 1} \frac{\sum_{r=1}^{R} (s_{Y_{1,t}^*} - \bar{s})^2}{\sum_{r=1}^{R} s_{Y_{1,t}^*}}$$

for $s = R - 1 \sum_{r=1}^{R} s_{Y_{1,t}^*}$ and $s_{Y_{1,t}^*} = (\hat{\epsilon}_{Y_{1,t}^*}^2 - (\hat{\epsilon}_{Y_{1,t}^*}^2 + (\hat{\gamma}_{Y_{1,t}^*} - \hat{Y}_{1,t}^*)^2)^2$. Table 2 records whether the Clark–West statistic leads to rejection based on the $N(0, 1)$ approximation to a one-sided test. For 51 of the 56 possible specifications of $Y_{1t}$, the out-of-sample evidence that $Y_{2,1,t}$ helps forecast $Y_{1t}$ is statistically significant at the 5% level for at least one of the elements of $Y_{1t}$.

One might think that perhaps the issue is that there may be more than 3 factors in $Y_{1t}$. We repeated the in-sample tests of whether $Y_{2}$ Granger-causes $Y_{1}$ for each of the 70 possible ways that a 4-dimensional vector $Y_{1t}$ could be chosen from $Y_{1}$, For 66 of these possibilities, the likelihood ratio test leads to rejection, and the 4 that are not rejected by this test turn out to be inconsistent with the $Y_{2}$ Granger-causality tests reported in the next subsection. If we let $Y_{1t}$ be a 5-dimensional vector, 52 of the 56 possibilities are rejected, and again the 4 that are not rejected here will be rejected by the tests below. Twenty-three of the 28 possible choices for a 6-dimensional factor vector are rejected. And even if we say that 7 of the 8 yields in $Y_{1}$ are themselves term-structure factors, for 5 of the 8 possible choices, we find that the one omitted yield Granger-causes the remaining 7.

Even if no single choice for the yields to include in $Y_{1t}$ is consistent with the data, is there some other linear combination of $Y_{1}$ that satisfies the Granger-causality restriction? One popular choice is to use the first 3 principal components of $Y_{1t}$ as the value for $Y_{1t}$, that is, use $Y_{1t} = (z_{1t}, z_{2t}, z_{3t})$ for $z_{r} = h_{r}Y_{1t}$ and $h_{r}$ the eigenvector associated with the ith largest eigenvalue of

$$T^{-1} \sum_{t=1}^{T} Y_{1t}Y_{1t}' = (20)$$

where elements of $\tilde{Y}_{1}$ are obtained by subtracting the mean of the corresponding elements of $Y_{1t}$. The first row of Table 3 reports $p$-values for tests that the first 3 principal components can be predicted from the last 5, both individually (first 3 columns) and as a group (last column). For example, we just fail to reject ($p = 0.061$) that $\alpha_{4} = \alpha_{5} = \cdots = \alpha_{9} = 0$ in the regression

$$z_{1t} = \alpha_{0} + \sum_{j=1}^{M} \alpha_{j}z_{j,t-1} + \varepsilon_{1t},$$

for $i \in \{1, 2, 3\}$ and $j \in \{4, 5, 6\}$.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
\begin{tabular}{c}
Specification of $Y_{1t}$
\end{tabular} & \begin{tabular}{c}
Out-of-sample improvement in MSE
\end{tabular} & \begin{tabular}{c}
1st
\end{tabular} & \begin{tabular}{c}
2nd
\end{tabular} & \begin{tabular}{c}
3rd
\end{tabular} \\
\hline
6m, 2y, 10y & 25% & 8% & 0% \\
6m, 2y, 3m & 25% & 9% & -1% \\
6m, 2y, 1y & 19% & 5% & 21% \\
6m, 2y, 3y & 10% & 6% & 4% \\
6m, 2y, 5y & 17% & 7% & 4% \\
6m, 2y, 7y & 15% & 7% & 3% \\
6m, 10y, 3m & -4% & 2% & -17% \\
6m, 10y, 1y & 4% & -1% & 16% \\
6m, 10y, 3y & 28% & 0% & 6% \\
6m, 10y, 5y & 29% & -2% & 2% \\
6m, 10y, 7y & 33% & -2% & 0% \\
6m, 3m, 1y & 39% & 16% & 18% \\
6m, 3m, 3y & 11% & -10% & 7% \\
6m, 3m, 5y & -2% & -18% & 6% \\
6m, 3m, 7y & -5% & -19% & 5% \\
6m, 1y, 3y & 5% & 16% & 4% \\
6m, 1y, 5y & -3% & 13% & 3% \\
6m, 1y, 7y & -4% & 13% & 2% \\
6m, 3y, 5y & 23% & 6% & 4% \\
6m, 3y, 7y & 18% & 6% & 3% \\
6m, 5y, 7y & 16% & 3% & 2% \\
2y, 10y, 3m & 7% & -1% & 27% \\
2y, 10y, 1y & 8% & 0% & 25% \\
2y, 10y, 3y & 5% & -3% & 1% \\
2y, 10y, 5y & 2% & -4% & 2% \\
2y, 10y, 7y & -3% & -2% & -1% \\
2y, 3m, 1y & 5% & 13% & 14% \\
2y, 3m, 3y & 4% & 6% & 3% \\
2y, 3m, 5y & 5% & 15% & 3% \\
2y, 3m, 7y & 6% & 17% & 2% \\
2y, 1y, 3y & 7% & 22% & 5% \\
2y, 1y, 5y & 8% & 25% & 4% \\
2y, 1y, 7y & 8% & 23% & 3% \\
2y, 3y, 5y & 7% & 4% & 3% \\
2y, 3y, 7y & 6% & 4% & 1% \\
2y, 5y, 7y & 8% & 2% & 1% \\
10y, 3m, 1y & 0% & 0% & 9% \\
10y, 3m, 3y & -1% & 36% & 6% \\
10y, 3m, 5y & -2% & 43% & 3% \\
10y, 3m, 7y & -2% & 53% & 0% \\
10y, 1y, 3y & 0% & 24% & 5% \\
10y, 1y, 5y & -2% & 23% & 1% \\
10y, 1y, 7y & -2% & 15% & 0% \\
10y, 3y, 5y & -3% & 0% & -1% \\
10y, 3y, 7y & -1% & 0% & 0% \\
10y, 5y, 7y & 1% & 4% & 3% \\
3m, 1y, 3y & -1% & 8% & 4% \\
3m, 1y, 5y & -7% & 5% & 4% \\
3m, 1y, 7y & -7% & 6% & 3% \\
3m, 3y, 5y & 28% & 5% & 3% \\
3m, 3y, 7y & 26% & 6% & 2% \\
3m, 5y, 7y & 30% & 4% & 2% \\
1y, 3y, 7y & 24% & 6% & 4% \\
1y, 3y, 7y & 21% & 6% & 3% \\
1y, 5y, 7y & 22% & 3% & 2% \\
3y, 5y, 7y & 6% & 3% & 2% \\
\hline
\end{tabular}
\caption{Out-of-sample Granger causality tests of null hypothesis that $Y_{2}$ does not Granger-cause $Y_{1}$ for alternative specifications for $Y_{1}$. Table entries report percent improvement in MSE for the equation that includes $Y_{2,1,t}$ over the equation that does not. Based on recursive regressions generating out-of-sample forecasts for 2002:M8–2007:M7.}
\end{table}
Table 3
In-sample Granger causality tests that last \( N - m \) principal components do not Granger-cause the first \( m \) for various values of \( m \). Table entries report \( p \)-values. The first 7 columns report predictability of \( z_t \), the \( j \)th principal component of \( Y_t \), on the basis of \( z_{t-1}, \ldots, z_{t-1} \), while the last column reports predictability of the vector \( (z_{t-1}, \ldots, z_{t-1})' \) on the basis of \( z_{t-1}, \ldots, z_{t-1} \). All regressions include \( (z_{t-1}, \ldots, z_{t-1})' \) and were estimated 1983:1:2002:M7.

<table>
<thead>
<tr>
<th>Number of principal components</th>
<th>Granger-causality tests</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m = 3 )</td>
<td></td>
<td>0.6099</td>
<td>0.1958</td>
<td>0.6889</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.687</td>
</tr>
<tr>
<td>( m = 4 )</td>
<td></td>
<td>0.0340</td>
<td>0.5386</td>
<td>0.5695</td>
<td>0.0294</td>
<td></td>
<td></td>
<td></td>
<td>0.0150</td>
</tr>
<tr>
<td>( m = 5 )</td>
<td></td>
<td>0.1493</td>
<td>0.4269</td>
<td>0.7404</td>
<td>0.5477</td>
<td>0.0214</td>
<td></td>
<td></td>
<td>0.0276</td>
</tr>
<tr>
<td>( m = 6 )</td>
<td></td>
<td>0.3783</td>
<td>0.7817</td>
<td>0.6129</td>
<td>0.3961</td>
<td>0.0537</td>
<td>0.0016</td>
<td></td>
<td>0.0000</td>
</tr>
<tr>
<td>( m = 7 )</td>
<td></td>
<td>0.1675</td>
<td>0.6911</td>
<td>0.2421</td>
<td>0.2816</td>
<td>0.0817</td>
<td>0.0030</td>
<td>0.5170</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

\( * \) Denoting rejection at the 5% level.

\( ** \) Denoting rejection at 1% level.

Table 4
Out-of-sample Granger causality test that last \( N - m \) principal components do not Granger-cause the first \( m \) for various values of \( m \). Table entries report percent improvement in MSE for the equation that includes last \( N - m \) principal components over the equation that does not. Table estimates represent out-of-sample improvement in MSE for the equation that includes \( z_{t-1}, \ldots, z_{t-1} \), over the equation that does not. The \( j \)th column reports predictability of \( z_t \), the \( j \)th principal component of \( Y_t \). Principal components estimated 1983:1:2002:M7 and evaluated using recursive regressions and out-of-sample forecasts for 2002:M8–2007:M7.

<table>
<thead>
<tr>
<th>Number of principal components</th>
<th>Out-of-sample improvement in MSE</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m = 3 )</td>
<td></td>
<td>7%</td>
<td>-4%</td>
<td>-3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m = 4 )</td>
<td></td>
<td>4%</td>
<td>-4%</td>
<td>-2%</td>
<td>-2%</td>
<td>-2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m = 5 )</td>
<td></td>
<td>6%</td>
<td>-3%</td>
<td>-2%</td>
<td>-1%</td>
<td>-1%</td>
<td>-8%</td>
<td>-5%</td>
<td>-1%</td>
</tr>
<tr>
<td>( m = 6 )</td>
<td></td>
<td>3%</td>
<td>0%</td>
<td>-2%</td>
<td>-2%</td>
<td>-5%</td>
<td>-5%</td>
<td>-5%</td>
<td>-2%</td>
</tr>
<tr>
<td>( m = 7 )</td>
<td></td>
<td>2%</td>
<td>0%</td>
<td>-1%</td>
<td>-1%</td>
<td>5%*</td>
<td>-2%</td>
<td>2%*</td>
<td></td>
</tr>
</tbody>
</table>

\( * \) Denoting Clark–West statistic leads to rejection of the null hypothesis of no improvement in the forecast at the 5% level.

\( ** \) Denoting rejection at 1% level.

(row 1, column 1) and likewise just fail to reject the joint hypothesis that \( \{z_{2t}, z_{2t}, z_{2t}\} \) cannot be predicted on the basis of \( \{z_{2t-1}, \ldots, z_{2t-1}\} \) (row 1, last column). Notwithstanding, these tests are quite close to rejection, and one might wonder whether 3 principal components may not be enough to capture the dynamics. But an interesting thing happens when we let \( Y_t \) be a \( (4 \times 1) \) vector corresponding to the first 4 principal components. As seen in the second row of Table 3, the evidence for statistical predictability is stronger when we use 4 principal components rather than 3. Indeed, we would also reject a specification using 5, 6, or even 7 principal components.

Table 4 investigates the predictability of principal components out of sample.\(^7\) While the contribution of \( \{z_{2t-1}, \ldots, z_{2t-1}\} \) is not quite statistically significantly helpful for forecasting \( z_{2t} \) within sample (first row and column of Table 3), it is statistically significantly helpful out of sample (first row and column of Table 4). Indeed, for all but one choice of the number of principal components to use in constructing \( Y_t \), there is at least one element of \( Y_t \) that can be forecast statistically significantly out of sample on the basis of \( Y_{2t-1} \).

What does the consistency with the data become even worse when we add more principal components? The assumption behind the ATSM was that, if we use enough principal components, we can capture the true factors, and whatever is left over is measurement or specification error, which was simply assumed to be white noise. But the feature of the data is that, even though the higher principal components are tiny, they are in fact still serially correlated. One can see this directly by looking at the vector autoregression for the elements of \( Y_{2t} \) alone.

\[
Y_{2t} = c_2 + \phi_{22} Y_{2t-1} + \xi_{2t}.
\]

Suppose we let \( Y_{2t} = (z_{m+1, t}, z_{m+2, t}, \ldots, z_{m})' \) be the smallest principal components and test whether \( \phi_{22} = 0 \), that is, test the null hypothesis that \( Y_{2t} \) is serially uncorrelated. This hypothesis turns out to be rejected at the 1% level for each choice of \( m = 3, 4, 5, 6, \) or 7. Moreover, cross-correlations between these smaller principal components are statistically significant, which explains why even though it may be hard to forecast \( \{z_{2t}, z_{2t}, z_{2t}\} \) using \( \{z_{4t-1}, z_{4t-1}, z_{4t-1}, z_{4t-1}, z_{4t-1}, z_{4t-1}, z_{4t-1}, z_{4t-1}\} \), it is in fact easier to forecast \( \{z_{1t}, z_{2t}, z_{3t}, z_{4t}\} \) using \( \{z_{5t-1}, z_{6t-1}, z_{7t-1}, z_{8t-1}\} \).

3.3. Granger-causality tests: \( Y_2 \)

We turn next to testable implications of (16), which embodies two sets of constraints. The first is that the \( m \) linear combinations of \( Y_t \) represented by \( Y_{2t} \) are sufficient to capture all the contemporaneous correlations. Specifically, if \( w_{2t} \) is any element of the \( N_r = (N - m) \) dimensional vector \( Y_{2t} \) and \( Y_{2t}^{(m)} \) denotes the remaining \( N_r - 1 \) elements of \( Y_{2t} \), then in the regression

\[
y_{2t} = c_0 + c_1 Y_{2t} + c_2 Y_{2t}^{(m)} + u_{2t}, \tag{21}
\]

we should find \( c_2 = 0 \). The first row of Table 5 reports in-sample \( p \)-values associated with the test of this null hypothesis when \( Y_{2t} \) is specified as the 6-month, 2-year, and 10-year yields. For \( y_{2t} \) the 3-month yield, we fail to reject the null hypothesis (\( p = 0.139 \)) that \( c_2 = 0 \). However, for each of the 4 other yields in \( Y_{2t} \) (namely, the 1-year, 3-year, 5-year, and 7-year), the null hypothesis is rejected at the 0.1% significance level, as reported in the remaining entries of the first row of Table 5. Subsequent rows of Table 5 report the analogous tests for every possible selection of 3 yields to include in \( Y_{2t} \). For every single choice, at least 4 of the resulting 5 elements in \( Y_{2t} \) are predictable, at a significance level less than 1%, by some of the other yields in \( Y_{2t} \); for both in-sample tests (Table 5) and out of sample (Table 6).

Nor can this problem obviously be solved by making \( Y_{2t} \) a higher-dimensional vector. For the \( \left( \begin{array}{c} 5 \\ 8 \end{array} \right) = 70 \) possible 4-dimensional vectors for \( Y_{2t} \), in every single case at least one of the elements of \( Y_{2t} \) is predictable at the 0.1% significance level by the other 3. For the \( \left( \begin{array}{c} 8 \\ 5 \end{array} \right) = 56 \) possible 5-dimensional vectors, all but 8

\( ^{7} \) Note that we keep \( h_t \) the same for each \( r \), that is, \( h_t \) is based on (20) for the original sample through \( T \), so that for each \( T + r \) we are talking about forecasting the same variable.
have at least one $y_{mt}$ for which the null hypothesis of no prediction is rejected at the 1% level. If we go to $m = 6$, of the 28 possible specifications of the 2-dimensional vector $Y_{2t}$, for 15 of them we find evidence at the 5% level that one is predicted by the other.

Note that when $Y_{1t}$ and $Y_{2t}$ consist of selected principal components, the elements are orthogonal by construction so the specification would necessarily pass the above test.

A separate implication of (16) is that, if we condition on the contemporaneous value of $Y_{1t}$, lagged values of $Y_{t-1}$ should be of no help in predicting the value of any element of $Y_{2t}$. That is, in the regression

$$y_{mt} = c_0 + c_1 y_{1t} + c_2 Y_{t-1} + u_{mt},$$

where $u_{mt}$ is the error term.
Table 7
In-sample tests of null hypothesis that lagged $Y_{t-1}$ does not help predict once $m$ contemporaneous principal components are included in the regression. The row $m$ and column $j$ entry reports $p$-value for predicting the $j$th principal component $z_j$ when the contemporaneous values of the first $m$ principal components are included.

<table>
<thead>
<tr>
<th>Number of principal components</th>
<th>Predictability tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4th</td>
</tr>
<tr>
<td>$m = 3$</td>
<td>0.000***</td>
</tr>
<tr>
<td>$m = 4$</td>
<td>0.000***</td>
</tr>
<tr>
<td>$m = 5$</td>
<td>0.000***</td>
</tr>
<tr>
<td>$m = 6$</td>
<td>0.000***</td>
</tr>
<tr>
<td>$m = 7$</td>
<td></td>
</tr>
</tbody>
</table>

* Denotes rejection at 5% level.
** Denotes rejection at 1% level.

Table 8
Out-of-sample tests of null hypothesis that lagged $Y_{t-1}$ does not help predict once $m$ contemporaneous principal components are included in the regression. The row $m$ and column $j$ entry reports the percent improvement in MSE for predicting the $j$th principal component $z_j$ when the contemporaneous values of the first $m$ principal components are included. Based on recursive regressions generating out-of-sample forecasts for 2002:M5–2007:M7.

<table>
<thead>
<tr>
<th>Number of principal components</th>
<th>Predictability tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4th</td>
</tr>
<tr>
<td>$m = 3$</td>
<td>75%</td>
</tr>
<tr>
<td>$m = 4$</td>
<td>11%</td>
</tr>
<tr>
<td>$m = 5$</td>
<td>41%</td>
</tr>
<tr>
<td>$m = 6$</td>
<td>68%</td>
</tr>
<tr>
<td>$m = 7$</td>
<td></td>
</tr>
</tbody>
</table>

* Indicates that the Clark–West statistic leads to rejection of the null hypothesis of no improvement in the forecast at the 5% level.
** Denotes rejection at 1% level.

The values of $\phi_1^m$ in (12) and $\phi_2^m$ in (16) are completely determined by the matrix $\rho$ and the sequence $\{b_n\}$, where the latter in turn can be calculated as functions of $\rho^0$ and $\delta_1$ using (2) and (7). The resulting value for $B_1$, along with the structural parameters $\Sigma$ and $\Sigma_e$, determine the variance–covariance matrix of the innovations in (12) and (16). Thus the likelihood function is fully specified by the structural parameters $\{c, \rho, c^0, \rho^0, \delta_1, \Sigma, \Sigma_e, \delta_0\}$. As discussed in Hamilton and Wu (2012b), some further normalization is necessary in order to be able to identify these structural parameters on the basis of observation of $\{Y_t\}_{t=1}^\infty$.

If we assume that $m$ linear combinations of $Y_t$ are observed without error, Jostlin et al. (2011) suggest that a natural normalization is to take the $(m \times 1)$ vector $F_t$ to be given by these particular linear combinations, $F_t = H_1Y_t$, for $H_1$ a known $(m \times N)$ matrix. For our base case specification in which $Y_t = (Y_{8t}, Y_{9t}, \ldots, Y_{24t}, Y_{36t}, Y_{60t}, Y_{84t}, Y_{120t})'$ and $Y_{1t} = (Y_{6t}, Y_{24t}, Y_{120t})'$, we would have

$$H_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$ 

Premultiplying (9) by $H_1$, and substituting the condition $F_t = H_1Y_t$ gives

$$H_1Y_t = H_1A + H_1B_1H_1Y_t,$$

requiring $H_1A = 0$ and $H_1B_1 = I_m$. These conditions turn out to imply a normalization similar to that of Jostlin et al. (2011) in which the $(m \times m)$ matrix $\rho_0$ is known up to its eigenvalues and the vector $c^0$ is a known function of those eigenvalues along with $\delta_0$ and $\Sigma$, as the following proposition demonstrates.

Proposition 1. Let $\xi = (\xi_1, \ldots, \xi_m)'$ denote a proposed vector of ordered eigenvalues of $\rho_0$. Let $t$ denote an $(m \times 1)$ vector of ones and $H_1$ a known $(m \times N)$ matrix. Define

$$y_n(x) = n^{-1} \sum_{j=0}^{n-1} x'$$

$$K(\xi) = \begin{bmatrix} Y_{n1}(\xi_1) & Y_{n2}(\xi_1) & \cdots & Y_{nm}(\xi_1) \\ Y_{n1}(\xi_2) & Y_{n2}(\xi_2) & \cdots & Y_{nm}(\xi_2) \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1}(\xi_m) & Y_{n2}(\xi_m) & \cdots & Y_{nm}(\xi_m) \end{bmatrix},$$

$$V(\xi) = \begin{bmatrix} \xi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \xi_m \end{bmatrix},$$

$$\rho(\xi) = [K(\xi)H_1']^{-1} [V(\xi)][K(\xi)H_1']$$

$$\delta_1 = [K(\xi)H_1']^{-1} t.$$ 

Then for $b_n$ constructed from (2) and (7) it is the case that

$$\begin{bmatrix} b_{n1} & \cdots & b_{nm} \end{bmatrix} H_1' = I_m.$$  \quad (22)
For given scalar $\delta_0$ and $(m \times m)$ matrix $\Sigma$ if we further define

$$
\zeta_n(\xi) = n^{-1} \left[ b_1' + 2b_2' + \cdots + (n-1)b_{n-1}' \right]
$$

$$
\psi_n(\xi, \delta_0, \Sigma) = \delta_0 - (2n-1)^{-1} \left[ b_1' \Sigma' b_1 + 2b_2' \Sigma' b_2 + \cdots + (n-1)^2 b_{n-1}' \Sigma' b_{n-1} \right]
$$

$$
c^Q_{(m \times 1)} = - \left( H_1 \left[ \begin{array}{c} \zeta_n(\xi) \\ \vdots \\ \zeta_n(\xi) \end{array} \right] \right)^{-1} \left( H_1 \psi_n(\xi, \delta_0, \Sigma) \right),
$$

(23)

then for $a_n$ constructed from (4) and (8), it is the case that

$$
H_1 \left[ \begin{array}{c} a_{n1} \\ \vdots \\ a_{nD} \end{array} \right] = \left[ \begin{array}{c} 0 \\ \vdots \\ 0 \end{array} \right].
$$

Suppose we assume that the factors are directly observable in the form of some known linear combination $Y_{1t} = H_1 Y_t$ and define those linear combinations observed with error to be $Y_{2t} = H_2 Y_t$ for $H_2$ a known $(N_c \times N)$ matrix with $N_c = N - m$. Then Proposition 1 establishes that the likelihood function of $\{Y_{1t}\}_{t=1}^T$ can be parameterized in terms of $(\xi, \rho, \Sigma, \delta_0)$. While the conventional approach to parameter estimation would be to choose these parameters so as to maximize the likelihood function directly, Hamilton and Wu (2012b) argue that there are substantial benefits from estimating by the minimum-chi-square procedure originally developed by Rothenberg (1973). The procedure is asymptotically equivalent to MLE but often substantially easier to implement. The approach is to first estimate the reduced-form parameters in (12) and (16) directly by ordinary least squares:

$$
\hat{\Pi}_{1t}' = \hat{A}_{1t}^* \hat{\phi}_{1t}^* = \left( \sum_{t=1}^T Y_{1t} x_{1t}' \right) \left( \sum_{t=1}^T x_{1t} x_{1t}' \right)^{-1} w_{1t} \cdot w_{1t} = \left( \begin{array}{c} 1 \\ Y_{1t-1}' \end{array} \right)
$$

$$
\hat{\Pi}_{1t} = T^{-1} \sum_{t=1}^T \left( Y_{1t} - \hat{\Pi}_{1t} x_{1t} \right) \left( Y_{1t} - \hat{\Pi}_{1t} x_{1t} \right)' \cdot w_{1t} = \left( \begin{array}{c} \hat{\Pi}_{1t} \\ \hat{\Pi}_{2t} \end{array} \right)
$$

$$
\hat{\Omega}_{2t} = T^{-1} \sum_{t=1}^T \left( Y_{2t} - \hat{\Pi}_{2t} x_{2t} \right) \left( Y_{2t} - \hat{\Pi}_{2t} x_{2t} \right)' \cdot w_{2t} = \left( \begin{array}{c} \hat{\Pi}_{2t} \\ \hat{\Pi}_{3t} \end{array} \right)
$$

The minimum-chi-square approach is to let these simple closed-form OLS formulas do the job of maximizing the unrestricted likelihood for $\{Y_t, \ldots, Y_T|Y_0\}$, and then find estimates of the structural parameters $(\xi, \rho, \Sigma, \Sigma, \delta_0)$ whose predicted values for these reduced-form coefficients are as close as possible to the OLS estimates. Closeness is defined in terms of minimizing a quadratic form with weighting matrix given by a consistent estimate of the information matrix:

$$
\hat{\delta}_{MCs} = \arg \min_{\delta} T \left[ \hat{\pi} - g(\theta) \right]' \hat{R} \left[ \hat{\pi} - g(\theta) \right].
$$

(24)

Here $\hat{\pi}$ is the vector of reduced-form parameters,

$$
\hat{\pi} = \left( \begin{array}{c} \text{vec}(\hat{\Pi}_{1t}') \\ \text{vec}(\hat{\Pi}_{2t}') \end{array} \right),
$$

$$\left( \begin{array}{c} \text{vec}(\hat{\Pi}_{1t}') \\ \text{vec}(\hat{\Pi}_{2t}') \end{array} \right)'.
$$

(25)

for $\text{vec}(\hat{\Pi}_{1t}')$ the $m(m+1) \times 1$ vector obtained by stacking columns of $\hat{\Pi}_{1t}$, and $\text{vec}(\hat{\Omega}_{2t}')$ the $m(m+1)/2 \times 1$ vector from stacking those elements in $\hat{\Omega}_{2t}$ that are on or below the principal diagonal. Also, $g(\theta)$ is the vector of predicted values for $\pi$ using the expressions in Section 3.1, while $\hat{R}$ is a matrix whose diagonal blocks are given by $\hat{\Omega}_{1t}^{-1} \otimes T^{-1} \sum_{t=1}^T x_{1t} x_{1t}'$, and $(1/2)D_m' (\hat{\Omega}_{1t}^{-1} \otimes \hat{\Omega}_{1t}^{-1})D_m$, $\hat{\Omega}_{2t}^{-1} \otimes T^{-1} \sum_{t=1}^T x_{2t} x_{2t}'$, and $(1/2)D_n' (\hat{\Omega}_{2t}^{-1} \otimes \hat{\Omega}_{2t}^{-1})D_n$, and whose other elements are all zero, and where $D_m$ denotes the $m^2 \times m(m+1)/2$ duplication matrix satisfying $D_m \text{ vec}(\Omega) = \text{ vec}(\Omega)$.

Note that since the information matrix is block diagonal with respect to the elements of $\hat{\Omega}_{2t}$, and since $\hat{\Omega}_{2t}$ are the only reduced-form parameters affected by the measurement error parameters $\Sigma_v$. MCSE for the latter can be obtained directly from the OLS estimates $\hat{\Omega}_{2t}$, namely $\hat{\Sigma}_v \hat{\Sigma}_v' = \hat{\Omega}_{2t}$, and this does not affect estimates of any other structural parameters. Moreover, this result still holds even when restrictions are imposed on $\Sigma_v$. For example, for the usual specification in which the measurement error is taken to be contemporaneously uncorrelated, the MCSE is obtained by setting diagonal elements of $\hat{\Sigma}_v$ equal to the square roots of the corresponding diagonal elements of $\hat{\Omega}_{2t}$, with off-diagonal elements of $\hat{\Sigma}_v$ set to zero, and again with no consequences for other parameter estimates.

Similarly, no matter what values might be chosen for the other parameters, as long as $B_1$ is invertible, from Eq. (13) we can always choose $\hat{c}$ so as to match $\hat{A}_1^*$ exactly, and from (14) we can choose $\hat{\rho}$ so as to match $\hat{\phi}_1^*$ exactly, so that the first block of $\hat{\pi}$ contributes zero to the objective function (24). Thus the numerical component of MCS estimation amounts to choosing $(\xi, \Sigma, \delta_0)$ so as to minimize

$$
T \left[ \hat{\pi}_2 - g(\theta) \right]' \hat{R}_2 \left[ \hat{\pi}_2 - g(\theta) \right]
$$

(26)

$$
\hat{\pi}_2 = \left( \begin{array}{c} \text{vec}(\hat{\Omega}_{1t}') \\ \text{vec}(\hat{\Pi}_{2t}') \end{array} \right),
$$

$$\left( \begin{array}{c} \text{vec}(\hat{\Omega}_{1t}') \\ \text{vec}(\hat{\Pi}_{2t}') \end{array} \right)'.
$$

$$\hat{\rho}_2 = \left[ \begin{array}{cc} (1/2)D_m' (\hat{\Omega}_{1t}^{-1} \otimes \hat{\Omega}_{1t}^{-1})D_m & 0 \\ 0 & \hat{\Omega}_{1t}^{-1} \otimes T^{-1} \sum_{t=1}^T x_{2t} x_{2t}' \end{array} \right].
$$

In addition to being asymptotically equivalent to and often easier to compute than the MLE, minimum-chi-square estimation has the further benefit that the optimized value for the objective function (26) has an asymptotic $\chi^2$ distribution with degrees of freedom given by the number of overidentifying restrictions. Hence an immediate by-product of the estimation is an evaluation of the validity of the kinds of restrictions considered in this section. There are $m(m+1)/2$ elements in $\hat{\Omega}_{1t}$ and $(N-m) (m+1)$ elements in $\hat{\Omega}_{2t}$, or 26 parameters in the unrestricted reduced form for the case when $m = 3$ and $N = 8$. On the other hand, there are $m$ elements in $\xi, m(m+1)/2$ elements in $\Sigma$, and 1 element in $\delta_0$, or 10 structural parameters for the above example. The model then imposes 16 overidentifying restrictions, or particular ways in which the parameters in regressions of the elements of $\Sigma_v$ on a constant and $Y_{1t}$ should be related to each other and related to the residual variance-covariance matrix for a VAR(1) for $Y_{1t}$ itself.

We first apply this procedure to our base-case specification in which $m = 3$ and $Y_{1t}$ is taken to be the 6-month, 2-year, and 10-year yields. The resulting $\chi^2(16)$ statistic is 633.58, leading to overwhelming rejection of the null hypothesis that the ATSM restrictions are consistent with the data. The procedure also

8 Joslin et al. (2011) derived a similar result for maximum likelihood estimation.
provides an immediate check on which elements of \( \hat{\pi}_2 \) are most at odds with the predictions implied by \( g_2(\hat{\theta}) \). The biggest positive contributions to (26) come from the constant terms \( A^*_0 \).

This claim might be surprising to many researchers, since it is often asserted that a standard ATSM does a good job of capturing the cross-section distribution of returns, precisely the claim being tested by the above \( \chi^2 \) test. The usual basis for the claim is the observation that 3 linear combinations of yields can account for an overwhelming fraction of the variances and covariances of yields. However, the high \( R^2 \) from such regressions only summarize the comovements between the variables as distinct from their individual average levels. The ATSM also has testable implications for the movements between the variables as distinct from their individual average levels. The ATSM also has testable implications for the movements between the variables as distinct from their individual average levels. The ATSM also has testable implications for the movements between the variables as distinct from their individual average levels. The ATSM also has testable implications for the movements between the variables as distinct from their individual average levels.

We can consider relaxing the feature of the ATSM by adding to each \( \alpha_m \) an unrestricted constant \( k_m \). This causes the parameter \( \delta_0 \) to be no longer identified, in effect replacing the original single parameter \( \delta_0 \) for purposes of describing the average values of the different yields with \( N = m \) new constants. The minimum value for (26) achieved by choice of \( \xi, \Sigma, k_{m+1}, \ldots, k_N \) turns out to be 132.75. Although this is a substantial improvement over the usual basis for the claim, the most promising specification was when \( Y_{1t} \) corresponds to the first 3 principal components, that is, \( Y_{1t} = H_1Y_t \) for rows of \( H_1 \) corresponding to the first three eigenvectors of (20), and \( Y_{2t} \) the remaining 5 principal components. When we calculate the MCS statistic (26) for the original specification, we arrive at a \( \chi^2(16) \) statistic of 650.47. Relaxing the constraint on the intercepts by introducing the \( k_{m+1}, \ldots, k_N \) parameters brings this down to \( \chi^2(12) = 145.05 \). Allowing for serial correlation in \( Y_{2t} \) yields a \( \chi^2(12) \) statistic of 13.48 (\( p = 0.335 \)), fully consistent with the data.

We conclude that representing the term structure factors by the first 3 principal components offers more promise of fitting the data than using any subset of \( m \) yields. However, it is necessary to acknowledge that the measurement or specification error is serially correlated. One furthermore needs to relax the predictions of the ATSM for the average levels of the various yields in order to describe accurately what is found in the data.

### 4. ATSM with observable macroeconomic factors

Up to this point we have been discussing models in which the only data being used are the yields themselves. There is a substantial literature beginning with Ang and Piazzesi (2003) that also incorporates directly observable macroeconomic variables such as output growth and inflation, collected in a vector \( f_t \). In our empirical investigation of these models, we will take \( f_t \) to be a \( (2 \times 1) \) vector whose first element is the monthly Chicago Fed National Activity Index and second element is the percentage change from the previous year in the implicit price deflator for personal consumption expenditures from the FRED database of the Federal Reserve Bank of St. Louis.

These observable macro factors \( f_t^o \) are then thought to supplement an \( (m \times 1) \) vector of conventional yield factors \( f_t^c \) in jointly determining the behavior of bond yields. The standard assumption is that the \( P \)-measure dynamics of the factors could be described with a VAR\(^{10} \):

\[
\begin{align*}
&f_t^c = c + \rho_1 f_{t-1}^c + \rho_2 f_{t-2}^c + \cdots + \rho_{oc} f_{t-k}^c + \epsilon_t^c, \\
&f_t^o = c_t + \rho_1 f_{t-1}^o + \rho_2 f_{t-2}^o + \cdots + \rho_{ot} f_{t-k}^o + \epsilon_t^o,
\end{align*}
\]

from which \( g_2(\hat{\theta}) = \text{vec} \left( \left[ A_2(\xi) B_2(\xi) \hat{\psi}_2(\xi) \right] \right) \) and (28) need only be minimized with respect to the 3 elements of \( \xi \).

\[^{9}\text{Implementing this turns out to be quite simple, since with } A_1^* \text{ unrestricted, } \\
\Sigma \text{ is unrestricted and the MCSE for } \Sigma \text{ satisfies } \Sigma \Sigma' = \hat{\Sigma}_j. \text{ Recall also } \\
b_1(\xi) = k_m. \text{ Moreover, given } \xi \text{ we can calculate } Y_{2t}(\xi) = Y_{2t} - B_1(\xi) Y_{1t} \text{ and } \\
[A_2(\xi) \hat{\psi}_2(\xi)] = \left( \sum_{t=1}^T Y_{2t}(\xi) \left[ \begin{array}{c} Y_{1t-1}^T \end{array} \right] \right) \left( \sum_{t=1}^T \left[ \begin{array}{c} Y_{1t-1} \end{array} \right] \right)^{-1}
\]

\[^{10}\text{The fact that only a single lag on } f_t^o \text{ is used is without loss of generality. If } \\
f_t^o \text{ is a latent vector, one could always stack a higher-order system for these } \\
latent variables into companion form, as we do below with the observed macro factors. However, if one wanted to take this interpretation of the dimension of } f_t^o, \text{ literally, one would want to impose corresponding additional restrictions on } \rho.\]
Defining $\mathbf{F}_k = (f_{1k}^o, f_{2k}^o, \ldots, f_{k-1k}^o)$, we can interpret (29)–(30) as an alternative formulation of (1) where $\mathbf{F}_k$ is now the $(2k + m) \times (2k + m)$ vector $\mathbf{F}_k = (\mathbf{F}_k^o, \mathbf{F}_k^l)$.

\[
\rho_{(2k+m) \times (2k+m)} = \begin{bmatrix}
\rho_{01} & \rho_{02} & \cdots & \rho_{0k-1} & \rho_{0k} & \rho_{00} \\
I_2 & 0 & \cdots & 0 & 0 & 0 \\
0 & I_2 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & I_2 & 0 & 0 \\
\rho_{11} & \rho_{12} & \cdots & \rho_{1k-1} & \rho_{1k} & \rho_{10} \\
\end{bmatrix}
\]

\[
\mathbf{c}_{(2k+m) \times 1} = (c_o, 0', \ldots, 0', c_i')'
\]

\[
\Sigma_{(2k+m) \times (2k+m)} = \begin{bmatrix}
\Sigma_0 & \cdots & 0 & \cdots & 0 \\
0 & \cdots & 0 & \cdots & 0 \\
0 & \cdots & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\Sigma_{10} & 0 & \cdots & \cdots & \Sigma_{k0} \\
\end{bmatrix}
\]

Elements of the $(m + 2k) \times 1$ vectors $\lambda$ and $\delta_1$ and of the $(m + 2k) \times (m + 2k)$ matrix $\mathbf{A}$ corresponding to zero blocks of $\Sigma$ are set to zero. We can then calculate predicted yields using (2) through (6) as before.

Among the choices to be made are the dimension of the latent vector $(m)$, number of lags to summarize macro dynamics $(k)$, and whether the macro factors and latent factors can be regarded as independent (as represented by the restrictions $\rho_{\delta\delta} = 0$, $\rho_{11} = \cdots = \rho_{kk} = 0$, and $\Sigma_{10} = 0$). Pericoli and Taboga (2008) conducted comprehensive investigations of this question through the arduous process of estimating assorted specifications subject to the full set of nonlinear restrictions imposed by the theory. Once again, however, it is possible to use Granger’s suggestion of choosing among the possible specifications on the basis of extremely simple tests of the underlying forecasting relations, as we now illustrate.

Suppose as in (10) that there is an $(m \times 1)$ vector of yields $\mathbf{Y}_t$, for which the predicted pricing relations hold exactly, and as in (11) that there is an $(N \times 1)$ vector $\mathbf{Y}_0$ priced with error. Then similar algebra to that used earlier produces the reduced form implied by the system:

\[
f_{1n}^o = A_{n0}^o + \phi_{1n}^o f_{1n-1}^o + \phi_{1n}^l Y_{1,t-1} + u_{nt}^o
\]

\[
f_{2n}^o = A_{n0}^o + \phi_{2n}^o f_{2n-1}^o + \phi_{2n}^l Y_{2,t-1} + u_{nt}^o
\]

\[
Y_{1t} = A_1^o + \phi_{1}^o f_{1t} + \phi_{1}^l Y_{1,t-1} + \psi_{1}^o f_{1t} + u_{1t}^o
\]

\[
Y_{2t} = A_2^o + \phi_{2}^o f_{2t} + \phi_{2}^l Y_{2,t-1} + u_{2t}^o.
\]

If the macro and finance factors are independent, then the coefficient $\phi_{2n}^l$ in (31) must be zero. Thus an immediate testable implication of independence of the macro and latent factors is whether the yields in $\mathbf{Y}_t$ Granger-cause the observed macro factors. Furthermore, the choice of $k$ ends up determining the number of lags of $f_{1n}^o, Y_{1t},$ and $Y_{2t}$ (dimensions of $\phi_{1n}^o, \phi_{1l}^o,$ and $\phi_{2n}^l$ in (31) through (33)). All of these can be tested by simple OLS without having to estimate the ATSM at all.

To illustrate this possibility, we focus on the choice in lag length between $k = 1$ or $k = 12$ and on whether one wants to model the latent factors and macro factors as independent. We further specify that $m = 3$ and that the 6-month, 2-year, and 10-year securities are priced without error. Row 2 of Table 9 indicates that we would reject the null hypothesis of independence under the maintained assumption of 12 lags, while row 3 indicates we would reject the null hypothesis that only 1 lag is needed under the maintained assumption of dependence.

### Table 9

<table>
<thead>
<tr>
<th>Lag length</th>
<th>Interaction</th>
<th>Likelihood ratio test</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 12$</td>
<td>Dependent</td>
<td>-</td>
<td>$-3094$</td>
<td>$-2038$</td>
</tr>
<tr>
<td>$k = 12$</td>
<td>Independent</td>
<td>$x^2 (6)$ = 15.9</td>
<td>$-3090$</td>
<td>$-2054$</td>
</tr>
<tr>
<td>$k = 1$</td>
<td>Dependent</td>
<td>$x^2 (220) = 442.3265$</td>
<td>$-3090$</td>
<td>$-2783$</td>
</tr>
<tr>
<td>$k = 1$</td>
<td>Independent</td>
<td>$x^2 (226) = 463.6942$</td>
<td>$-3080$</td>
<td>$-2794$</td>
</tr>
</tbody>
</table>

Denotes rejection at 1% level.

### 5. Conclusion

A number of previous researchers have discussed related shortcomings of ATSM. Cochrane and Piazzesi (2009) documented that the linear combinations that describe the contemporaneous correlations among yields are different from those that are most helpful for forecasting. Collin-Dufresne and Goldstein (2002) found that lagged volatilities as well as lagged levels of yields contribute to forecasts, while Ludvigson and Ng (2011), Cooper and Priestly (2009), and Joslin et al. (2010) concluded that macro variables have useful forecasting information beyond that contained in current yields. Duffee (2011b) suggested that these results could be explained by near-cancellation of the forecasting and risk-pricing implications of certain factors, causing these factors to be hidden from any collection of contemporaneous yields and yet still useful for forecasting future yields.

However, the results in our paper go beyond any of these claims. We find that for $Y_{1t}$ a collection of m yields or principal components and $Y_{2t}$ the remaining yields or components, the data consistently reject the hypothesis that $Y_t$ does not Granger-cause $Y_{1t}$, regardless of how large one makes $m$, and further reject the hypothesis that the residuals from a regression of $Y_{2t}$ on $Y_{1t}$ are serially uncorrelated. These results could not be attributed to hidden or omitted factors in the sense of Duffee (2011b) or Collin-Dufresne and Goldstein (2002), since our explanatory variables are direct functions of the yields themselves. Instead we find that the data speak conclusively that the specification or measurement error in the system must have its own important dynamic structure.

As noted by Duffee (2011a,b), the specification error could be broadly attributed to factors such as bid/ask spreads, preferred habitats of particular investors, interpolation errors, and liquidity.
premia. None of these factors would a priori be expected to be white noise, and it should not be surprising that we find the measurement error terms in these models to be quite predictable. Furthermore, it is not a defense to argue that this serial correlation can be ignored because the errors themselves are small—this form of model misspecification makes conventional standard errors unreliable and invalidates standard hypothesis tests about any parameters of the system.

In this paper, we suggested one approach to dealing with these problems, which is to postulate as a primitive that the specification errors have their own mean and serial dependence structure, and estimate these separately from the parameters of the core ATSM. We illustrated estimation of a system of this form that seems to be consistent with the data. A more satisfactory approach would be to try to understand the features of these specification errors in a more structural way, for example, trying to model liquidity effects directly. This seems a particularly important task if one’s goal is to understand the behavior of the term structure during the financial crisis in the fall of 2008, for which Gürkaynak and Wright (2012) showed that even the most basic arbitrage relations appeared to break down.

Apart from these issues, our paper illustrates that many of the key underlying assumptions of ATSM are trivially easy to test. Clive Granger’s perennial question of whether the model’s specification is consistent with basic forecasting relations in the data seems a particularly helpful guide for research using ATSM.

Appendix

Proof of Proposition 1. Observe that

\[
b_n = n^{-1} \left[ I_m + (\rho^1)^0 + (\rho^2)^0 + \cdots + (\rho^N)^0 \right] \cdot 1
\]

Furthermore, for \( \alpha_n \) satisfying (4) and (8) and \( \sigma^2 \) satisfying (23),

\[
H_1 \left[ \begin{array}{c} \alpha_n^0 \\ \vdots \\ \alpha_n^N \end{array} \right] = H_1 \left[ \begin{array}{c} \psi_n^0 (\xi, \delta_0, \Sigma) \\ \vdots \\ \psi_n^N (\xi, \delta_0, \Sigma) \end{array} \right] + H_1 \left[ \begin{array}{c} \alpha_n^0 (\xi) \\ \vdots \\ \alpha_n^N (\xi) \end{array} \right] \cdot \sigma^2
\]

References


