Testable Implications of Affine Term Structure Models*

James D. Hamilton†
Department of Economics
University of California, San Diego

Jing Cynthia Wu‡
Booth School of Business
University of Chicago

October 26, 2010
Revised: March 5, 2013

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.

*We thank Jonathan Wright and anonymous referees for helpful comments on an earlier draft of this paper.
†jhamilton@ucsd.edu
‡Cynthia.Wu@chicagobooth.edu

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, Pan and Singleton, 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, Swanson and Wu, 2006), inferring market expectations of inflation (Christensen, Lopez and Rudebusch, 2010), and evaluating the effects of the extraordinary central bank interventions during the financial crisis (Christensen, Lopez and Rudebusch, 2009; Smith, 2010; Hamilton and Wu, 2012a). Gürkaynak 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, Diebold and Rudebusch (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 ATSM 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 χ^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

[1] On the other hand, Duffee (2011a) found that the ATSM cross-section restrictions don't and shouldn't help with forecasting.

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 all of the possible formulations investigated here.[2] We further demonstrate that the predictions of these models for the average levels of different interest rates are inconsistent with the observed data. We find that a specification in which (1) the term structure factors are measured by the first three principal components of the set of observed yields , (2) predictions for average levels of interest rates are relaxed, and (3) measurement error is serially correlated, can be reconciled with the observed time series behavior of interest rates. We illustrate how Granger-causality tests can also be used to determine the specification of complicated macro-finance term structure models. Such tests suggest that a strong premium should be placed on parsimony.

2 Affine term structure models.

Let P_{nt} denote the price at time t of a pure-discount bond that is certain to be worth \$1 at time $t+n$. A broad class of finance models posit that $P_{nt} = E_t(M_{t+1}P_{n-1,t+1})$ for some pricing kernel M_{t+1}. Affine term structure models suppose that the price P_{nt} depends on a possibly unobserved $(m \times 1)$ vector of factors F_t that follows a Gaussian first-order vector autoregression,

$$F_{t+1} = c + \rho F_t + \Sigma u_{t+1} \tag{1}$$

with u_t an i.i.d. sequence of $N(0, I_m)$ vectors. The second component of ATSM is the assumption that the pricing kernel is characterized by $M_{t+1} = \exp\left(-r_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t' u_{t+1}\right)$ for r_t the risk-free one-period interest rate and λ_t an $(m \times 1)$ vector that characterizes investors' attitudes toward risk; $\lambda_t = 0$ would correspond to risk neutrality. Both this risk-pricing vector and the risk-free rate are postulated to be affine functions of the vector of factors: $\lambda_t = \lambda + \Lambda F_t$ and $r_t = \delta_0 + \delta_1' F_t$. The risk-free rate r_t is simply the negative of the log of the price of a one-period bond,

$$r_t = \log(P_{0t}/P_{1t}) = \log(1) - \log(P_{1t}) = -p_{1t},$$

[2]Duffee (2011a) has also noted the substantial serial correlation of measurement errors.

for $p_{nt} = \log(P_{nt})$. After a little algebra (e.g., Ang and Piazzesi, 2003), the above equations imply that

$$p_{nt} = \overline{a}_n + \overline{b}_n' F_t$$

where the values of $\overline{b}_n$ and $\overline{a}_n$ can be calculated recursively from

$$\overline{b}_n' = \overline{b}_{n-1}' \rho^Q - \delta_1' \tag{2}$$

$$\rho^Q = \rho - \Sigma\Lambda \tag{3}$$

$$\overline{a}_n = \overline{a}_{n-1} + \overline{b}_{n-1}' c^Q + (1/2)\overline{b}_{n-1}' \Sigma\Sigma' \overline{b}_{n-1} - \delta_0 \tag{4}$$

$$c^Q = c - \Sigma\lambda \tag{5}$$

starting from $\overline{b}_1 = -\delta_1$ and $\overline{a}_1 = -\delta_0$. The implied yield on an n-period bond, $y_{nt} = -n^{-1} p_{nt}$, is then characterized by

$$y_{nt} = a_n + b_n' F_t \tag{6}$$

$$b_n = -n^{-1} \overline{b}_n \tag{7}$$

$$a_n = -n^{-1} \overline{a}_n. \tag{8}$$

Suppose we observe a set of N different yields, $Y_t = (y_{n_1,t}, y_{n_2,t}, ..., y_{n_N,t})'$, and collect (6) into a vector system

$$Y_t = A + BF_t \tag{9}$$

for A an $(N \times 1)$ vector whose ith element is a_{n_i} and B an $(N \times m)$ matrix whose ith row is b_{n_i}'. If $m < N$, then the model (9) is instantly refuted, because it implies that a regression of any one of the yields on m 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,

$$Y_{1t} = A_1 + B_1 F_t \tag{10}$$

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, Singleton and Zhu, 2011). The

remaining $N_e = N - m$ yields are assumed to be priced with error,

$$Y_{2t} = A_2 + B_2 F_t + u_{2t} \tag{11}$$

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. $N(0, \Sigma_e \Sigma_e')$ for Σ_e 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),

$$Y_{1,t+1} = A_1^* + \phi_{11}^* Y_{1t} + u_{1,t+1} \tag{12}$$

$$A_1^* = A_1 + B_1 c - B_1 \rho B_1^{-1} A_1 \tag{13}$$

$$\phi_{11}^* = B_1 \rho B_1^{-1} \tag{14}$$

$$u_{1,t+1} = B_1 \Sigma u_{t+1}. \tag{15}$$

Similar operations on (11) produce

$$Y_{2t} = A_2^* + \phi_{21}^* Y_{1t} + u_{2t} \tag{16}$$

$$A_2^* = A_2 - B_2 B_1^{-1} A_1 \tag{17}$$

$$\phi_{21}^* = B_2 B_1^{-1}, \tag{18}$$

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 $\{Y_t, Y_{t-1}, ...\}$, and u_{2t} in (16) is uncorrelated with $\{Y_{t-1}, Y_{t-2}, ...\}$. 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 $(Y_{1t}', Y_{2t}')'$ 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.

3.2 Granger-causality tests: Y_1.

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

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

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 by 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 to 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 ($Y_{1t} = (y_{6t}, y_{24t}, y_{120t})'$), while the other yields are priced with error ($Y_{2t} = (y_{3t}, y_{12t}, y_{36t}, y_{60t}, y_{84t})'$). 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 ϕ_{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 ϕ_{12}^* are zero ($Y_{2,t-1}$ does not predict $y_{24,t}$ or $y_{120,t}$) fail to reject with p-values of 0.198 and 0.204. A likelihood ratio test with Sims' small-sample correction[4] of the

[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 $\hat{u}_{1t}$ denote the vector of OLS residuals from estimation of (19) over $t = 1, ..., T$ and $\hat{\Omega}_1 = T^{-1} \sum_{t=1}^{T} \hat{u}_{1t} \hat{u}_{1t}'$. Let $\tilde{u}_{1t}$ denote the vector of OLS residuals when $Y_{2,t-1}$ is dropped from the equation with $\hat{\Omega}_0 = T^{-1} \sum_{t=1}^{T} \tilde{u}_{1t} \tilde{u}_{1t}'$. Then as in Hamilton (1994), equation [11.1.34], $(T - N - 1)(\log \left|\hat{\Omega}_0\right| - \log \left|\hat{\Omega}_1\right|)$ is approximately $\chi^2(m(N-m))$ for N the dimension of Y_t and m the dimension of Y_{1t}. All system-wide

null hypothesis that all 15 elements of ϕ_{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_{2t} is priced with error, it should contain no information about the factors beyond that contained in Y_{1t}, and therefore should not help to predict $Y_{1,t+1}$. If some maturities are more helpful than others for forecasting, those are the ones we'd want to include in Y_{1t} 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_{1t}. 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 $\hat{y}_{i,T+1}$ and associated forecast error $\hat{\varepsilon}_{i,T+1}$. We also estimated the restricted regressions with $\phi_{12}^* = 0$ to calculate a restricted forecast $\hat{y}_{i,T+1}^*$ and error $\hat{\varepsilon}_{i,T+1}^*$. We then increased the sample size by one to generate $\hat{y}_{i,T+2}$ and $\hat{y}_{i,T+2}^*$, 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,

$$\frac{R^{-1}\sum_{r=1}^{R}\left[\left(\hat{\varepsilon}_{i,T+r}^*\right)^2 - \left(\hat{\varepsilon}_{i,T+r}\right)^2\right]}{R^{-1}\sum_{r=1}^{R}\left[\left(\hat{\varepsilon}_{i,T+r}^*\right)^2\right]}$$

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 $Y_{2,t-1}$ leads to a 25% out-of-sample improvement in forecasting the 6-month yield and an 8% improvement for the 2-year yield for $Y_{1t} = (y_{6t}, y_{24t}, y_{120t})'$.

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 $Y_{2,t-1}$ 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

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.

Normal, seems to be reasonably well approximated by the $N(0,1)$ distribution,

$$C = \frac{\sqrt{R}\overline{s}}{\sqrt{R^{-1}\sum_{r=1}^{R}(s_{T+r} - \overline{s})^2}}$$

for $\overline{s} = R^{-1}\sum_{r=1}^{R} s_{T+r}$ and $s_{T+r} = \left(\hat{\varepsilon}^*_{i,T+r}\right)^2 - \left(\hat{\varepsilon}_{i,T+r}\right)^2 + \left(\hat{y}^*_{i,T+r} - \hat{y}_{i,T+r}\right)^2$. Table 2 records whether the Clark-West statistic leads to rejection based on the $N(0,1)$ approximation to a one-sided test.[5] For 51 out of the 56 possible specifications of Y_{1t}, the out-of-sample evidence that $Y_{2,t-1}$ helps forecast Y_{1t} is statistically significant at the 5% level for at least one of the elements of Y_{1t}.[6]

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 $\binom{8}{4} = 70$ possible ways that a 4-dimensional vector Y_{1t} could be chosen from Y_t. 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_t 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_t that satisfies the Granger-causality restriction? One popular choice is to use the first 3 principal components of Y_t as the value for Y_{1t}, that is, use $Y_{1t} = (z_{1t}, z_{2t}, z_{3t})'$ for $z_{it} = h_i' Y_t$ and h_i the eigenvector associated with the ith largest eigenvalue of

$$T^{-1}\sum_{t=1}^{T} \tilde{Y}_t \tilde{Y}_t' \tag{20}$$

where elements of $\tilde{Y}_t$ are obtained by subtracting the mean of the corresponding elements of Y_t. 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).

[5] That is, * indicates a value of C above 1.645 and ** a value above 2.33.

[6] One might note that the biggest out-of-sample improvements come from yields of 1-year maturity or less. We attribute this to the fact that over the post-sample evaluation period (2002:M8 to 2007:M7), short rates exhibited a dramatic swing down and back up while long rates remained fairly flat– there is simply more for the regression to forecast with short rates than long rates on this subsample.

For example, we just fail to reject ($p = 0.061$) that $\alpha_4 = \alpha_5 = \cdots = \alpha_8 = 0$ in the regression

$$z_{1t} = \alpha_0 + \sum_{j=1}^{8} \alpha_j z_{j,t-1} + \varepsilon_{1t},$$

(row 1, column 1) and likewise just fail to reject the joint hypothesis that $\{z_{1t}, z_{2t}, z_{3t}\}$ cannot be predicted on the basis of $\{z_{j,t-1}\}_{j=4}^{8}$ (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_{1t} be a (4×1) 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'd 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 $\{z_{4,t-1}, ..., z_{8,t-1}\}$ is not quite statistically significantly helpful for forecasting z_{1t} 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_{1t}, there is at least one element of Y_{1t} that can be forecast statistically significantly out of sample on the basis of $Y_{2,t-1}$.

Why 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 in 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,

$$Y_{2t} = c_2 + \phi_{22} Y_{2,t-1} + \varepsilon_{2t}.$$

Suppose we let $Y_{2t} = (z_{m+1,t}, z_{m+2,t}, ..., z_{Nt})'$ 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 $\{z_{1t}, z_{2t}, z_{3t}\}$ using $\{z_{4,t-1}, z_{5,t-1}, z_{6,t-1}, z_{7,t-1}, z_{8,t-1}\}$, it is in fact easier to forecast $\{z_{1t}, z_{2t}, z_{3t}, z_{4t}\}$ using

[7]Note that we keep h_i the same for each r, that is, h_i is based on (20) for the original sample through T, so that for each $T + r$ we are talking about forecasting the same variable.

$\{z_{5,t-1}, z_{6,t-1}, z_{7,t-1}, z_{8,t-1}\}$.

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_{1t} are sufficient to capture all the contemporaneous correlations. Specifically, if y_{nt} is any element of the $N_e = (N - m)$ dimensional vector Y_{2t} and $Y_{2t}^{(n)}$ denotes the remaining $N_e - 1$ elements of Y_{2t}, then in the regression

$$y_{nt} = c_0 + c_1' Y_{1t} + c_2' Y_{2t}^{(n)} + u_{nt}, \tag{21}$$

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_{1t} is specified as the 6-month, 2-year, and 10-year yields. For y_{nt} 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_{1t}. 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_{1t} a higher-dimensional vector. For the $\binom{8}{4} = 70$ possible 4-dimensional vectors for Y_{1t}, 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 $\binom{8}{5} = 56$ possible 5-dimensional vectors, all but 8 have at least one y_{nt} 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 that 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

$$y_{nt} = c_0 + c_1' Y_{1t} + c_2' Y_{t-1} + u_{nt},$$

we should find that the 8 elements of c_2 are all zero if y_{nt} is any element of Y_{2t}. For each of the 56 possible choices for the 3-dimensional vector Y_{1t}, this hypothesis ends up being rejected at the 1% level for each of the implied 5 elements of Y_{2t} on the basis of both the in-sample F

test and the out-of-sample Clark-West test.

Using a higher-dimensional Y_{1t} or principal components does not solve this problem. For example, let z_{jt} denote the jth principal component and consider the regression

$$z_{jt} = c_0 + \sum_{i=1}^{m} c_{1i} z_{it} + c_2' Y_{t-1} + u_{jt}$$

for some $j > m$. The first row of Table 7 shows that, for $m = 3$, we strongly reject the hypothesis that $c_2 = 0$ for each $j = 4, 5, 6, 7, 8$. Subsequent rows show that the same is true for any choice of m. Table 8 confirms that the statistical contribution of Y_{t-1} to a forecast of any of the smaller principal components is statistically significant out of sample as well.

Our conclusion from this and the preceding subsection is that the assumption that there exists a readily observed factor of any dimension that captures all the predictability of Y_t is not consistent with the behavior of these data. At a minimum, a data-coherent specification requires the assumption that the measurement or specification error must be serially correlated.

3.4 Tests of predicted values for nonzero coefficients.

Up to this point we have been testing the large blocks of zero restrictions imposed by equations (12) and (16) relative to an unrestricted VAR. We now consider the particular values predicted by an ATSM for the nonzero elements in these two equations. Duffee (2011a) used mean-squared-error comparisons to conclude that these nonlinear restrictions are typically rejected statistically. Here we use the minimum-chi-square approach to test overidentifying restrictions developed by Hamilton and Wu (2012b). First we will develop some new extensions of those methods appropriate for the case in which the factors F_t are treated as directly observed in the sense that the value of B_1 in (10) is known a priori; the alternative case of latent factors (that is, when B_1 must be estimated) is discussed in Hamilton and Wu (2012b). Note that the tests described in Sections 3.2-3.3 are perfectly valid regardless of whether the factors are treated as latent or observed.

The values of ϕ_{11}^* in (12) and ϕ_{21}^* in (16) are completely determined by the matrix ρ and the sequence $\{b_n\}$, where the latter in turn can be calculated as functions of ρ^Q and δ_1 using (2) and (7) . The resulting value for B_1, along with the structural parameters Σ and Σ_e, determine the variance-covariance matrix of the innovations in (12) and (16). The sequence $\{b_n\}$ and values of Σ, c^Q, δ_0, c and ρ can be used to calculate the constants A_1^* and A_2^* in (12) and (16). Thus the likelihood function is fully specified by the structural parameters $\{c, \rho, c^Q, \rho^Q, \delta_1, \Sigma, \Sigma_e, \delta_0\}$.

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 $\{Y_t\}_{t=1}^T$.

If we assume that m linear combinations of Y_t are observed without error, Joslin, Singleton and Zhu (2011) suggest that a natural normalization is to take the $(m \times 1)$ vector F_t to be given by these particular linear combinations, $F_t = H_1 Y_t$, for H_1 a known $(m \times N)$ matrix. For our base case specification in which $Y_t = (y_{3t}, y_{6t}, y_{12,t}y_{24,t}, y_{36,t}, y_{60,t}, y_{84,t}, y_{120,t})'$ and $Y_{1t} = (y_{6t}, y_{24,t}, y_{120,t})'$ we would have

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

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

$$H_1 Y_t = H_1 A + H_1 B H_1 Y_t,$$

requiring $H_1 A = 0$ and $H_1 B = I_m$. These conditions turn out to imply a normalization similar to that of Joslin, Singleton and Zhu (2011) in which the $(m \times m)$ matrix ρ^Q is known up to its eigenvalues and the vector c^Q is a known function of those eigenvalues along with δ_0 and Σ, as the following proposition demonstrates.

Proposition 1. *Let $\xi = (\xi_1, ..., \xi_m)'$ denote a proposed vector of ordered eigenvalues of ρ^Q. Let ι denote an $(m \times 1)$ vector of ones and H_1 a known $(m \times N)$ matrix. Define*

$$\underset{(1\times 1)}{\gamma_n(x)} = n^{-1} \sum_{j=0}^{n-1} x^j$$

$$\underset{(m\times N)}{K(\xi)} = \begin{bmatrix} \gamma_{n_1}(\xi_1) & \gamma_{n_2}(\xi_1) & \cdots & \gamma_{n_N}(\xi_1) \\ \gamma_{n_1}(\xi_2) & \gamma_{n_2}(\xi_2) & \cdots & \gamma_{n_N}(\xi_2) \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{n_1}(\xi_m) & \gamma_{n_2}(\xi_m) & \cdots & \gamma_{n_N}(\xi_m) \end{bmatrix}$$

$$\underset{(m\times m)}{V(\xi)} = \begin{bmatrix} \xi_1 & \cdots & 0 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & \xi_m \end{bmatrix}$$

$$\underset{(m\times m)}{\rho^{Q\prime}} = [K(\xi)H_1']^{-1} [V(\xi)] [K(\xi)H_1']$$

$$\underset{(m\times 1)}{\delta_1} = [K(\xi)H_1']^{-1}\iota.$$

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

$$\left[\begin{array}{ccc} b_{n_1} & \cdots & b_{n_N} \end{array}\right] H_1' = I_m. \tag{22}$$

For given scalar δ_0 and $(m \times m)$ matrix Σ, if we further define

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

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

$$\underset{(m\times 1)}{c^Q} = -\left(H_1 \left[\begin{array}{c} \zeta_{n_1}(\xi) \\ \vdots \\ \zeta_{n_N}(\xi) \end{array}\right]\right)^{-1} H_1 \left[\begin{array}{c} \psi_{n_1}(\xi, \delta_0, \Sigma) \\ \vdots \\ \psi_{n_N}(\xi, \delta_0, \Sigma) \end{array}\right], \tag{23}$$

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

$$H_1 \left[\begin{array}{c} a_{n_1} \\ \vdots \\ a_{n_N} \end{array}\right] = \left[\begin{array}{c} 0 \\ \vdots \\ 0 \end{array}\right].$$

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 $(N_e \times N)$ matrix with $N_e = N - m$. Then Proposition 1 establishes that the likelihood function of $\{Y_t\}_{t=1}^T$ can be parameterized in terms of $\{c, \rho, \xi, \Sigma, \Sigma_e, \delta_0\}$. 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:

$$\hat{\Pi}_1^{*\prime} = \left[\begin{array}{cc} \hat{A}_1^* & \hat{\phi}_{11}^* \end{array}\right] = \left(\sum_{t=1}^T Y_{1t} x_{1t}'\right)\left(\sum_{t=1}^T x_{1t} x_{1t}'\right)^{-1}$$

$$x_{1t}' = \left[\begin{array}{cc} 1 & Y_{1,t-1}' \end{array}\right]$$

$$\hat{\Omega}_1^* = T^{-1}\sum_{t=1}^{T}\left(Y_{1t} - \hat{\Pi}_1^{*\prime}x_{1t}\right)\left(Y_{1t} - \hat{\Pi}_1^{*\prime}x_{1t}\right)'$$

$$\hat{\Pi}_2^{*\prime} = \left[\begin{array}{cc} \hat{A}_2^* & \hat{\phi}_{21}^* \end{array}\right] = \left(\sum_{t=1}^{T} Y_{2t}x_{2t}'\right)\left(\sum_{t=1}^{T} x_{2t}x_{2t}'\right)^{-1}$$

$$x_{2t}' = \left[\begin{array}{cc} 1 & Y_{1t}' \end{array}\right]$$

$$\hat{\Omega}_2^* = T^{-1}\sum_{t=1}^{T}\left(Y_{2t} - \hat{\Pi}_2^{*\prime}x_{2t}\right)\left(Y_{2t} - \hat{\Pi}_2^{*\prime}x_{2t}\right)'.$$

The minimum-chi-square approach is to let these simple closed-form OLS formulas do the job of maximizing the unrestricted likelihood for $\{Y_1, ..., Y_T|Y_0\}$, and then find estimates of the structural parameters $\{c, \rho, \xi, \Sigma, \Sigma_e, \delta_0\}$ 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:

$$\hat{\theta}_{MCS} = \arg\min_{\theta} T\left[\hat{\pi} - g(\theta)\right]'\hat{R}\left[\hat{\pi} - g(\theta)\right]. \tag{24}$$

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

$$\hat{\pi} = \left(\left[\text{vec}(\hat{\Pi}_1^*)\right]', \left[\text{vech}(\hat{\Omega}_1^*)\right]', \left[\text{vec}(\hat{\Pi}_2^*)\right]', \left[\text{vech}(\hat{\Omega}_2^*)\right]'\right)', \tag{25}$$

for $\text{vec}(\hat{\Pi}_1^*)$ the $m(m+1) \times 1$ vector obtained by stacking columns of $\hat{\Pi}_1^*$, and $\text{vech}(\hat{\Omega}_1^*)$ the $m(m+1)/2 \times 1$ vector from stacking those elements in $\hat{\Omega}_1^*$ that are on or below the principal diagonal. Also, $g(\theta)$ is the vector of predicted values for π using the expressions in Section 3.1, while $\hat{R}$ is a matrix whose diagonal blocks are given by $\hat{\Omega}_1^{*-1} \otimes T^{-1}\sum_{t=1}^{T} x_{1t}x_{1t}'$, $(1/2)D_m'(\hat{\Omega}_1^{*-1} \otimes \hat{\Omega}_1^{*-1})D_m$, $\hat{\Omega}_2^{*-1} \otimes T^{-1}\sum_{t=1}^{T} x_{2t}x_{2t}'$, and $(1/2)D_{N_e}'(\hat{\Omega}_2^{*-1} \otimes \hat{\Omega}_2^{*-1})D_{N_e}$, and whose other elements are all zero, and where D_m denotes the $m^2 \times m(m+1)/2$ duplication matrix satisfying $D_m \text{vech}(\Omega) = \text{vec}(\Omega)$.

Note that since the information matrix is block diagonal with respect to the elements of Ω_2^*, and since Ω_2^* are the only reduced-form parameters affected by the measurement error parameters Σ_e, MCSE for the latter can be obtained directly from the OLS estimates $\hat{\Omega}_2^*$, namely $\hat{\Sigma}_e\hat{\Sigma}_e' = \hat{\Omega}_2^*$, and this does not affect estimates of any other structural parameters. Moreover, this result still holds even when restrictions are imposed on Σ_e. 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 $\hat{\Sigma}_e$ equal to the square

roots of the corresponding diagonal elements of $\hat{\Omega}_2^*$, with off-diagonal elements of $\hat{\Sigma}_e$ 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 equation (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 $\hat{\phi}_{11}^*$ exactly, so that the first block of $\hat{\pi}$ contributes zero to the objective function (24).[8] Thus the numerical component of MCS estimation amounts to choosing $\{\xi, \Sigma, \delta_0\}$ so as to minimize

$$T\left[\hat{\pi}_2 - g_2(\theta)\right]' \hat{R}_2 \left[\hat{\pi}_2 - g_2(\theta)\right] \tag{26}$$

$$\hat{\pi}_2 = \left(\left[\text{vech}(\hat{\Omega}_1^*)\right]', \left[\text{vec}(\hat{\Pi}_2^*)\right]' \right)'$$

$$\hat{R}_2 = \begin{bmatrix} (1/2)D_m'(\hat{\Omega}_1^{*-1} \otimes \hat{\Omega}_1^{*-1})D_m & 0 \\ 0 & \hat{\Omega}_2^{*-1} \otimes T^{-1}\sum_{t=1}^{T} x_{2t}x_{2t}' \end{bmatrix}.$$

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 χ^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 $\hat{\Omega}_1^*$ and $(N-m)(m+1)$ elements in $\hat{\Pi}_2^*$, 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 ξ, $m(m+1)/2$ elements in Σ, and 1 element in δ_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 Y_{2t} 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 provides an immediate check on which elements of $\hat{\pi}_2$ are most at odds with the predictions implied by $g_2(\hat{\theta}_2)$. The biggest positive contributions to (26) come from the constant terms $\hat{A}_2^*$.

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

[8] Joslin, Singleton and Zhu (2011) derived a similar result for maximum likelihood estimation.

claim being tested by the above χ^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 latter, which we have just seen are inconsistent with the values observed in the data.

We can consider relaxing this feature of the ATSM by adding to each a_n an unrestricted constant k_n. This causes the parameter δ_0 to be no longer identified, in effect replacing the original single parameter δ_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 $\{\xi, \Sigma, k_{m+1}, ..., k_N\}$ turns out to be 132.75. Although this is a substantial improvement over the original specification, it is still grossly inconsistent with a $\chi^2(12)$ distribution.

Although the MCS χ^2 statistic is not directly testing the separate zero restrictions that we investigated earlier, some of those restrictions are maintained auxiliary assumptions that can influence the outcome of the χ^2 test. In particular, we saw above that there is very strong evidence that the error term in the Y_{2t} regression is serially correlated. We now investigate MCS estimation of an ATSM when this restriction is relaxed.

Suppose that (16) holds, with ϕ^*_{21} given by the structural parameters $B_2 B_1^{-1}$ but A^*_2 unrestricted and the error term correlated with lagged yields:

$$u_{2t} = \psi_2 Y_{t-1} + \varepsilon_{2t}. \tag{27}$$

Substituting (27) into (16) results in

$$Y_{2t} = A^\dagger_2 + B_2 B_1^{-1} Y_{1t} + \psi_2 Y_{t-1} + \varepsilon_{2t}$$

for which the corresponding unrestricted reduced form is

$$Y_{2t} = A^\dagger_2 + \psi^\dagger_1 Y_{1t} + \psi^\dagger_2 Y_{t-1} + \varepsilon_{2t}$$

whose unrestricted estimates are again easily obtained by OLS. We then choose $\{\xi, \Sigma, A^\dagger_2, \psi^\dagger_2\}$

so as to minimize[9]

$$T\left[\hat{\pi}_2^\dagger - g_2^\dagger(\theta)\right]' \hat{R}_2^\dagger \left[\hat{\pi}_2^\dagger - g_2^\dagger(\theta)\right] \tag{28}$$

where $\hat{\pi}_2^\dagger = \left(\left[\text{vech}(\hat{\Omega}_1^*)\right]', \left[\text{vec}(\hat{\Pi}_2^\dagger)\right]'\right)'$, $x_{2t}^{\dagger\prime} = \begin{bmatrix} 1 & Y_{1t}' & Y_{t-1}' \end{bmatrix}$, and

$$\hat{R}_2^\dagger = \begin{bmatrix} (1/2)D_m'(\hat{\Omega}_1^{*-1} \otimes \hat{\Omega}_1^{*-1})D_m & 0 \\ 0 & \hat{\Omega}_2^{\dagger-1} \otimes T^{-1}\sum_{t=1}^T x_{2t}^\dagger x_{2t}^{\dagger\prime} \end{bmatrix}$$

$$\hat{\Pi}_2^{\dagger\prime} = \left(\sum_{t=1}^T Y_{2t} x_{2t}^{\dagger\prime}\right)\left(\sum_{t=1}^T x_{2t}^\dagger x_{2t}^{\dagger\prime}\right)^{-1}$$

$$\hat{\Omega}_2^\dagger = T^{-1}\sum_{t=1}^T \left(Y_{2t} - \hat{\Pi}_2^{\dagger\prime} x_{2t}^\dagger\right)\left(Y_{2t} - \hat{\Pi}_2^{\dagger\prime} x_{2t}^\dagger\right)'.$$

For this case, there are $m(m+1)/2+N_e(1+m+N) = 66$ unrestricted reduced-form parameters and $m+m(m+1)/2+N_e(N+1) = 54$ structural parameters for 12 overidentifying restrictions. The $\chi^2(12)$ statistic turns out to be 78.52, which still leads to strong rejection.

One could relax additional restrictions to try to arrive at a specification that is not rejected. However, even if a specification were found that is consistent with the observed value for Π_2, the model would still have to contend with rejection of the many separate zero restrictions documented above. Based on those earlier tests, the most promising specification was when Y_{1t} corresponds to the first 3 principal components, that is, $Y_{1t} = H_1 Y_t$ 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 $k_{m+1}, ..., k_N$ parameters brings this down to $\chi^2(12) = 145.05$. Allowing for serial correlation in u_{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.

[9]Implementing this turns out to be quite simple, since with $A_2^\dagger$ unrestricted, Σ is unrestricted and the MCSE for Σ satisfies $\hat{\Sigma}\hat{\Sigma}' = \hat{\Omega}_1^*$. Recall also $B_1(\xi) = I_m$. Moreover, given ξ we can calculate $\tilde{Y}_{2t}(\xi) = Y_{2t} - B_2(\xi)Y_{1t}$ and $\begin{bmatrix} \tilde{A}_2(\xi) & \tilde{\psi}_2(\xi) \end{bmatrix} = \left(\sum_{t=1}^T \tilde{Y}_{2t}(\xi)\begin{bmatrix} 1 & Y_{t-1}' \end{bmatrix}\right)\left(\sum_{t=1}^T \begin{bmatrix} 1 \\ Y_{t-1} \end{bmatrix}\begin{bmatrix} 1 & Y_{t-1}' \end{bmatrix}\right)^{-1}$ from which $g_2(\xi) = \text{vec}\left(\begin{bmatrix} \tilde{A}_2(\xi) & B_2(\xi) & \tilde{\psi}_2(\xi) \end{bmatrix}'\right)$ and (28) need only be minimized with respect to the 3 elements of ξ.

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^o. In our empirical investigation of these models, we will take f_t^o to be a (2×1) 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^ℓ 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]

$$f_t^o = c_o + \rho_{o1} f_{t-1}^o + \rho_{o2} f_{t-2}^o + \cdots + \rho_{ok} f_{t-k}^o + \rho_{o\ell} f_{t-1}^\ell + \Sigma_{oo} u_t^o \tag{29}$$

$$f_t^\ell = c_\ell + \rho_{\ell 1} f_{t-1}^o + \rho_{\ell 2} f_{t-2}^o + \cdots + \rho_{\ell k} f_{t-k}^o + \rho_{\ell\ell} f_{t-1}^\ell + \Sigma_{\ell o} u_t^o + \Sigma_{\ell\ell} u_t^\ell. \tag{30}$$

Defining $F_t^o = (f_t^{o\prime}, f_{t-1}^{o\prime}, ..., f_{t-k+1}^{o\prime})'$, we can interpret (29)-(30) as an alternative formulation of (1) where F_t is now the $(2k+m)$ vector $F_t = (F_t^{o\prime}, f_t^{\ell\prime})'$,

$$\underset{(2k+m)\times(2k+m)}{\rho} = \begin{bmatrix} \rho_{o1} & \rho_{o2} & \cdots & \rho_{o,k-1} & \rho_{ok} & \rho_{o\ell} \\ I_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & I_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & I_2 & 0 & 0 \\ \rho_{\ell 1} & \rho_{\ell 2} & \cdots & \rho_{\ell,k-1} & \rho_{\ell k} & \rho_{\ell\ell} \end{bmatrix}$$

$$\underset{(2k+m)\times 1}{c} = (c_o', 0', ..., 0', c_\ell')'$$

[10] The fact that only a single lag on f_t^ℓ is used is without loss of generality. If f_t^ℓ 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^ℓ literally, one would want to impose corresponding additional restrictions on ρ.

$$\underset{(2k+m)\times(2k+m)}{\Sigma} = \begin{bmatrix} \Sigma_{oo} & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Sigma_{\ell o} & 0 & 0 & \cdots & \Sigma_{\ell\ell} \end{bmatrix}.$$

Elements of the $(m + 2k) \times 1$ vectors λ and δ_1 and of the $(m + 2k) \times (m + 2k)$ matrix Λ corresponding to zero blocks of Σ 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 $\rho_{o\ell} = 0$, $\rho_{\ell 1} = \cdots = \rho_{\ell k} = 0$, and $\Sigma_{\ell o} = 0$). 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 Y_{1t} for which the predicted pricing relations hold exactly, and as in (11) that there is an $(N_e \times 1)$ vector Y_{2t} priced with error. Then similar algebra to that used earlier produces the reduced form implied by the system:

$$\underset{(2\times 1)}{f_t^o} = A_m^* + \underset{(2\times 2k)}{\phi_{oo}^*} F_{t-1}^o + \underset{(2\times m)}{\phi_{o1}^*} Y_{1,t-1} + u_{ot}^* \tag{31}$$

$$\underset{(m\times 1)}{Y_{1t}} = A_1^* + \underset{(m\times 2k)}{\phi_{1o}^*} F_{t-1}^o + \underset{(m\times m)}{\phi_{11}^*} Y_{1,t-1} + \underset{(m\times 2)}{\psi_{1o}^*} f_t^o + u_{1t}^* \tag{32}$$

$$\underset{(N_e\times 1)}{Y_{2t}} = A_2^* + \underset{(N_e\times 2k)}{\phi_{2o}^*} F_t^o + \underset{(N_e\times m)}{\phi_{21}^*} Y_{1t} + u_{2t}^*. \tag{33}$$

If the macro and finance factors are independent, then the coefficient ϕ_{o1}^* in (31) must be zero. Thus an immediate testable implication of independence of the macro and latent factors is whether the yields in Y_{1t} Granger-cause the observed macro factors. Furthermore, the choice of k ends up determining the number of lags of f_{t-j}^o that are helpful for forecasting f_t^o, Y_{1t}, and Y_{2t} (dimensions of ϕ_{oo}^*, ϕ_{1o}^*, and ϕ_{2o}^* 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.

Despite the superior in-sample fit, the least restrictive specification in row 1 of Table 9 is richly parameterized, with 28 to 30 regression coefficients estimated per equation. While the model selection criterion[11] suggested by Akaike reaches the same conclusion as the in-sample F test, the Schwarz criterion favors the most parsimonious specification with $k = 1$ and independence of the macro and latent factors. Table 10 reinforces this conclusion from Schwarz, finding that the out-of-sample, one-month-ahead forecast of yields generated by the $k = 1$ specifications always beat $k = 12$. On the other hand, a specification that allows dependence between the macro and latent factors usually dominates the independent specification in terms of out-of-sample performance. These results suggest that a parsimonious 1-lag specification that still allows for interaction between the factors might be preferred.

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, Priebsch and Singleton (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_2 does not Granger-cause Y_1, 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

[11] See for example Lütkepohl (1993), p. 202.

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) and Duffee (2011b), 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.

6 Appendix

Proof of Proposition 1. Observe that

$$b_n = n^{-1}\left[I_m + \left(\rho^{Q\prime}\right) + \left(\rho^{Q\prime}\right)^2 + \cdots + \left(\rho^{Q\prime}\right)^{n-1}\right]\delta_1$$

$$\left(\rho^{Q\prime}\right)^s = \left[K(\xi)H_1'\right]^{-1}\left[V(\xi)\right]^s\left[K(\xi)H_1'\right]$$

$$\left[V(\xi)\right]^s = \begin{bmatrix} \xi_1^s & \cdots & 0 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & \xi_m^s \end{bmatrix}$$

$$\begin{aligned} b_n &= n^{-1}\left\{\left[K(\xi)H_1'\right]^{-1}\left[I_m + V(\xi) + V(\xi)^2 + \cdots V(\xi)^{n-1}\right]\left[K(\xi)H_1'\right\}\left[K(\xi)H_1'\right]^{-1}\iota \right. \\ &= \left[K(\xi)H_1'\right]^{-1}\begin{bmatrix} \gamma_n(\xi_1) & \cdots & 0 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & \gamma_n(\xi_m) \end{bmatrix}\iota \\ &= \left[K(\xi)H_1'\right]^{-1}\begin{bmatrix} \gamma_n(\xi_1) \\ \vdots \\ \gamma_n(\xi_m) \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \begin{bmatrix} b_{n_1} & \cdots & b_{n_N} \end{bmatrix} H_1' &= \left[K(\xi)H_1'\right]^{-1}\begin{bmatrix} \gamma_{n_1}(\xi_1) & \cdots & \gamma_{n_N}(\xi_1) \\ \vdots & & \vdots \\ \gamma_{n_1}(\xi_m) & \cdots & \gamma_{n_N}(\xi_m) \end{bmatrix} H_1' \\ &= \left[K(\xi)H_1'\right]^{-1} K(\xi)H_1' \\ &= I_m. \end{aligned}$$

Furthermore, for a_n satisfying (4) and (8) and c^Q satisfying (23),

$$\begin{aligned}
H_1 \begin{bmatrix} a_{n_1} \\ \vdots \\ a_{n_N} \end{bmatrix} &= H_1 \begin{bmatrix} \psi_{n_1}(\xi, \delta_0, \Sigma) \\ \vdots \\ \psi_{n_N}(\xi, \delta_0, \Sigma) \end{bmatrix} + H_1 \begin{bmatrix} \zeta_{n_1}(\xi) \\ \vdots \\ \zeta_{n_N}(\xi) \end{bmatrix} c^Q \\
&= H_1 \begin{bmatrix} \psi_{n_1}(\xi, \delta_0, \Sigma) \\ \vdots \\ \psi_{n_N}(\xi, \delta_0, \Sigma) \end{bmatrix} - H_1 \begin{bmatrix} \zeta_{n_1}(\xi) \\ \vdots \\ \zeta_{n_N}(\xi) \end{bmatrix} \times \\
&\quad \left(H_1 \begin{bmatrix} \zeta_{n_1}(\xi) \\ \vdots \\ \zeta_{n_N}(\xi) \end{bmatrix} \right)^{-1} H_1 \begin{bmatrix} \psi_{n_1}(\xi, \delta_0, \Sigma) \\ \vdots \\ \psi_{n_N}(\xi, \delta_0, \Sigma) \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.
\end{aligned}$$

References

Ang, Andrew and Monika Piazzesi, "A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables," *Journal of Monetary Economics*, 2003, *50*, 745–787.

Bauer, Michael D., "Term Premia and the News," 2011. Federal Reserve Bank of San Francisco, Working paper.

Beechey, Meredith J. and Jonathan H. Wright, "The High-Frequency Impact of News On Long-Term Yields and Forward Rates: Is It Real?," *Journal of Monetary Economics*, 2009, *56*, 535–544.

Chen, Ren-Raw and Louis Scott, "Maximum Likelihood Estimation for a Multifactor Equilibrium Model of the Term Structure of Interest Rates," *The Journal of Fixed Income*, 1993, *3*, 14–31.

Christensen, Jens H. E., Francis X. Diebold, and Glenn D. Rudebusch, "The Affine Arbitrage-Free Class of Nelson-Siegel Term Structure Models," *Journal of Econometrics*, 2011, *164* (1), 4–20.

_ , Jose A. Lopez, and Glenn D. Rudebusch, "Do Central Bank Liquidity Facilities Affect Interbank Lending Rates?," 2009. Working paper 2009-13, Federal Reserve Bank of San Francisco.

_ , _ , and _ , "Inflation Expectations and Risk Premiums in an Arbitrage-Free Model of Nominal and Real Bond Yields," *Journal of Money, Credit, and Banking*, 2010, *42*, 143–178.

Clark, Todd E. and Kenneth D. West, "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics*, 2007, *138*, 291–311.

Cochrane, John H. and Monika Piazzesi, "Decomposing the Yield Curve," 2009. AFA 2010 Atlanta Meetings Paper.

Collin-Dufresne, Pierre and Robert S. Goldstein, "Do Bonds Span the Fixed Income Markets? Theory and Evidence for Unspanned Stochastic Volatility," *Journal of Finance*, 2002, *57* (4), 1685–1730.

Cooper, Ilan and Richard Priestly, "Time-Varying Risk Premiums and the Output Gap," *Review of Financial Studies*, 2009, *22*, 2801–2833.

Dai, Qiang and Kenneth J. Singleton, "Specification Analysis of Affine Term Structure Models," *The Journal of Finance*, 2000, *55*, 1943–1978.

Duffee, Gregory R., "Term Premia and Interest Rate Forecasts in Affine Models," *The Journal of Finance*, 2002, *57*, 405–443.

_ , "Forecasting with the Term Structure: The Role of No-Arbitrage Restrictions," 2011. Working Paper, Johns Hopkins University.

_ , "Information in (and Not in) the Term Structure," *Review of Financial Studies*, 2011, *24* (9), 2895–2934.

Duffie, D., J. Pan, and K. Singleton, "Transform Analysis and Asset Pricing for Affine Jump-Diffusions," *Econometrica*, 2000, *68* (6), 1343–1376.

Granger, Clive W. J., "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, 1969, *37*, 424–438.

_ , "Testing for Causality: A Personal Viewpoint," *Journal of Economic Dynamics and Control*, 1980, *2*, 329–352.

Gürkaynak, Refet S. and Jonathan H. Wright, "Macroeconomics and the Term Structure," *Journal of Economic Literature*, 2012, *50*, 331–367.

_ **, Brian Sack, and Jonathan H. Wright**, "The U.S. Treasury Yield Curve: 1961 to the Present," *Journal of Monetary Economics*, November 2007, *54* (8), 2291–2304.

Hamilton, James D., *Time Series Analysis*, Princeton, New Jersey: Princeton University Press, 1994.

_ **and Jing Cynthia Wu**, "The Effectiveness of Alternative Monetary Policy Tools in a Zero Lower Bound Environment," *Journal of Money, Credit, and Banking*, 2012, *44 (s1)*, 3–46.

_ **and** _ , "Identification and Estimation of Gaussian Affine Term Structure Models," *Journal of Econometrics*, 2012, *168*, 315–331.

Hong, Yongmiao and Haitao Li, "Nonparametric Specification Testing for Continuous-Time Models with Applications to Term Structure of Interest Rates," *Review of Financial Studies*, 2005, *18*, 37–84.

Joslin, Scott, Kenneth J. Singleton, and Haoxiang Zhu, "A New Perspective On Gaussian Dynamic Term Structure Models," *Review of Financial Studies*, 2011, *27*, 926–970.

_ **, Marcel Priebsch, and Kenneth J. Singleton**, "Risk Premium Accounting in Dynamic Term Structure Models with Unspanned Macro Risks," 2010. Working paper, Stanford.

Ludvigson, Sidney C. and Serena Ng, "A Factor Analysis of Bond Risk Premia," in A. Ulah and D. Giles, eds., *Handbook of Empirical Economics and Finance*, Chapman and Hall, 2011, pp. 313–372.

Lütkepohl, Helmut, *Introduction to Multiple Time Series Analysis*, Berlin: Springer-Verlag, 1993.

Pericoli, Marcello and Marco Taboga, "Canonical Term-Structure Models with Observable Factors and the Dynamics of Bond Risk Premia," *Journal of Money, Credit and Banking*, 2008, *40*, 1471–1488.

Rothenberg, Thomas J., *Efficient Estimation with A Priori Information*, Yale University Press, 1973.

Rudebusch, Glenn D., "Macro-Finance Models of Interest Rates and the Economy," *The Manchester School (Supplement)*, 2010, pp. 25–52.

_ **and Tao Wu**, "A Macro-Finance Model of the Term Structure, Monetary Policy and the Economy," *The Economic Journal*, 2008, *118*, 906–926.

_ **, Eric T. Swanson, and Tao Wu**, "The Bond Yield 'Conundrum' From a Macro-Finance Perspective," *Monetary and Economic Studies (Special Edition)*, 2006, pp. 83–128.

Smith, Josephine M., "The Term Structure of Money Market Spreads During the Financial Crisis." PhD dissertation, Stanford University 2010.

Specification of Y_{1t}	Granger-causality tests			
	1^{st}	2^{nd}	3^{rd}	system
6m,2y,10y	0.006**	0.198	0.204	0.000**
6m,2y,3m	0.000**	0.219	0.002**	0.000**
6m,2y,1y	0.005**	0.223	0.021*	0.000**
6m,2y,3y	0.065	0.235	0.155	0.000**
6m,2y,5y	0.049*	0.242	0.086	0.000**
6m,2y,7y	0.020*	0.211	0.110	0.000**
6m,10y,3m	0.000**	0.522	0.004**	0.000**
6m,10y,1y	0.010**	0.205	0.022*	0.000**
6m,10y,3y	0.001**	0.215	0.116	0.000**
6m,10y,5y	0.000**	0.252	0.164	0.000**
6m,10y,7y	0.000**	0.232	0.206	0.000**
6m,3m,1y	0.003**	0.002**	0.021*	0.000**
6m,3m,3y	0.000**	0.003**	0.141	0.000**
6m,3m,5y	0.000**	0.004**	0.151	0.000**
6m,3m,7y	0.000**	0.004**	0.252	0.000**
6m,1y,3y	0.010**	0.022*	0.128	0.000**
6m,1y,5y	0.015*	0.024*	0.081	0.000**
6m,1y,7y	0.012*	0.023*	0.110	0.000**
6m,3y,5y	0.012*	0.118	0.082	0.001**
6m,3y,7y	0.004**	0.111	0.110	0.000**
6m,5y,7y	0.000**	0.124	0.121	0.000**
2y,10y,3m	0.183	0.243	0.006**	0.002**
2y,10y,1y	0.176	0.203	0.025*	0.000**
2y,10y,3y	0.216	0.296	0.249	0.000**
2y,10y,5y	0.602	0.314	0.391	0.002**
2y,10y,7y	0.346	0.241	0.263	0.001**
2y,3m,1y	0.182	0.004**	0.018*	0.000**
2y,3m,3y	0.188	0.080	0.112	0.000**
2y,3m,5y	0.193	0.060	0.080	0.001**
2y,3m,7y	0.183	0.023*	0.121	0.002**
2y,1y,3y	0.192	0.075	0.141	0.000**
2y,1y,5y	0.205	0.086	0.090	0.000**
2y,1y,7y	0.181	0.049*	0.112	0.000**
2y,3y,5y	0.176	0.109	0.081	0.000**
2y,3y,7y	0.165	0.159	0.126	0.000**
2y,5y,7y	0.278	0.179	0.134	0.000**
10y,3m,1y	0.289	0.009**	0.017*	0.018*
10y,3m,3y	0.259	0.001**	0.119	0.001**
10y,3m,5y	0.301	0.000**	0.184	0.000**
10y,3m,7y	0.264	0.000**	0.243	0.000**
10y,1y,3y	0.219	0.011*	0.128	0.000**
10y,1y,5y	0.263	0.010**	0.217	0.000**
10y,1y,7y	0.232	0.011*	0.217	0.000**
10y,3y,5y	0.282	0.357	0.332	0.015*
10y,3y,7y	0.230	0.179	0.235	0.000**
10y,5y,7y	0.222	0.110	0.161	0.000**
3m,1y,3y	0.010**	0.017*	0.106	0.000**
3m,1y,5y	0.018*	0.018*	0.090	0.019*
3m,1y,7y	0.013*	0.018*	0.142	0.049*
3m,3y,5y	0.014*	0.106	0.081	0.002**
3m,3y,7y	0.004**	0.109	0.124	0.000**
3m,5y,7y	0.000**	0.140	0.137	0.000**
1y,3y,5y	0.043*	0.108	0.082	0.000**
1y,3y,7y	0.020*	0.111	0.110	0.000**
1y,5y,7y	0.008**	0.145	0.124	0.000**
3y,5y,7y	0.188	0.164	0.129	0.004**

Table 1: In-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 p-values, with * denoting rejection at the 5% level and ** denoting rejection at the 1% level. First three columns report p-value for predictability of the ith element of Y_1, while last column tests predictability of the full vector Y_1. Regressions estimated 1983:M1-2002:M7.

Specification of Y_{1t}	Out-of-sample improvement in MSE		
	1^{st}	2^{nd}	3^{rd}
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%	17%**	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,5y	24%**	6%	4%
1y,3y,7y	21%**	6%*	3%
1y,5y,7y	22%**	3%	2%
3y,5y,7y	6%*	3%	2%

Table 2: 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 equation that includes $Y_{2,t-1}$ over equation that does not. Asterisk (*) denotes Clark-West statistic leads to rejection of the null hypothesis of no improvement in the forecast at the 5% level, while ** denotes rejection at 1% level. Based on recursive regressions generating out-of-sample forecasts for 2002:M8-2007:M7.

Number of	Granger-causality tests							
principal components	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	system
$m = 3$	0.0609	0.1958	0.6889	–	–	–	–	0.0687
$m = 4$	0.0340*	0.5368	0.5669	0.0294*	–	–	–	0.0150*
$m = 5$	0.1493	0.4269	0.7404	0.5477	0.0214*	–	–	0.0276*
$m = 6$	0.3783	0.7817	0.6129	0.3961	0.0537	0.0016**	–	0.0089**
$m = 7$	0.1675	0.6911	0.4241	0.2816	0.0817	0.0030**	0.5170	0.0050*

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, with * denoting rejection at the 5% level and ** denoting rejection at the 1% level. The first 7 columns report predictability of z_{jt}, the jth principal component of Y_t, on the basis of $z_{m+1,t-1}, ..., z_{N,t-1}$, while the last column reports predictability of the vector $(z_{1t}, ..., z_{mt})'$ on the basis of $z_{m+1,t-1}, ..., z_{N,t-1}$. All regressions include $(z_{1,t-1}, ..., z_{m,t-1})'$ and were estimated 1983:M1-2002:M7.

Number of	Out-of-sample improvement in MSE						
principal components	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}
$m = 3$	7%*	-4%	-3%	–	–	–	–
$m = 4$	4%*	-2%	-2%	-2%	–	–	–
$m = 5$	6%*	-3%	-2%	-1%	-8%	–	–
$m = 6$	3%	0%	-2%	-2%	-5%	-5%	–
$m = 7$	2%	0%	-1%	-1%	5%*	-2%	2%*

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 equation that includes last $N - m$ principal components over equation that does not. Asterisk (*) denotes Clark-West statistic leads to rejection of the null hypothesis of no improvement in the forecast at the 5% level, while ** denotes rejection at 1% level. Table estimates represent out-of-sample improvement in MSE for equation that includes $z_{m+1,t-1}, ..., z_{N,t-1}$ over equation that does not. Asterisk (*) denotes statistically significant contribution at the 5% level, and ** denotes significant at 1% level. The jth column reports predictability of z_{jt}, the jth principal component of Y_t. Principal components estimated 1983:M1-2002:M7 and evaluated using recursive regressions and out-of-sample forecasts for 2002:M8-2007:M7.

Specification of Y_{1t}	Ability to predict each element of Y_{2t}				
	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}
6m,2y,10y	0.139	0.000**	0.000**	0.000**	0.000**
6m,2y,3m	0.000**	0.000**	0.000**	0.000**	0.000**
6m,2y,1y	0.000**	0.001**	0.000**	0.000**	0.000**
6m,2y,3y	0.000**	0.038*	0.000**	0.000**	0.000**
6m,2y,5y	0.000**	0.076	0.000**	0.587	0.000**
6m,2y,7y	0.000**	0.284	0.000**	0.000**	0.000**
6m,10y,3m	0.000**	0.000**	0.000**	0.000**	0.000**
6m,10y,1y	0.000**	0.001**	0.000**	0.000**	0.000**
6m,10y,3y	0.000**	0.001**	0.000**	0.000**	0.000**
6m,10y,5y	0.000**	0.000**	0.000**	0.000**	0.000**
6m,10y,7y	0.000**	0.000**	0.000**	0.000**	0.000**
6m,3m,1y	0.000**	0.000**	0.000**	0.000**	0.000**
6m,3m,3y	0.000**	0.000**	0.000**	0.000**	0.000**
6m,3m,5y	0.000**	0.000**	0.000**	0.000**	0.000**
6m,3m,7y	0.000**	0.000**	0.000**	0.000**	0.000**
6m,1y,3y	0.000**	0.000**	0.001**	0.000**	0.000**
6m,1y,5y	0.000**	0.000**	0.001**	0.000**	0.000**
6m,1y,7y	0.000**	0.000**	0.001**	0.000**	0.000**
6m,3y,5y	0.000**	0.000**	0.000**	0.000**	0.000**
6m,3y,7y	0.000**	0.000**	0.002**	0.000**	0.002**
6m,5y,7y	0.000**	0.000**	0.000**	0.000**	0.000**
2y,10y,3m	0.000**	0.000**	0.000**	0.000**	0.000**
2y,10y,1y	0.000**	0.000**	0.000**	0.000**	0.000**
2y,10y,3y	0.000**	0.000**	0.000**	0.000**	0.000**
2y,10y,5y	0.000**	0.000**	0.000**	0.000**	0.000**
2y,10y,7y	0.000**	0.000**	0.000**	0.000**	0.000**
2y,3m,1y	0.009**	0.000**	0.000**	0.000**	0.000**
2y,3m,3y	0.000**	0.000**	0.000**	0.000**	0.000**
2y,3m,5y	0.000**	0.000**	0.000**	0.007**	0.000**
2y,3m,7y	0.000**	0.000**	0.000**	0.000**	0.000**
2y,1y,3y	0.000**	0.000**	0.000**	0.000**	0.000**
2y,1y,5y	0.000**	0.000**	0.000**	0.001**	0.000**
2y,1y,7y	0.000**	0.015*	0.000**	0.000**	0.000**
2y,3y,5y	0.000**	0.000**	0.000**	0.000**	0.000**
2y,3y,7y	0.000**	0.000**	0.000**	0.000**	0.448
2y,5y,7y	0.000**	0.000**	0.000**	0.000**	0.000**
10y,3m,1y	0.064	0.000**	0.000**	0.000**	0.000**
10y,3m,3y	0.000**	0.000**	0.000**	0.000**	0.000**
10y,3m,5y	0.000**	0.000**	0.000**	0.000**	0.000**
10y,3m,7y	0.000**	0.000**	0.000**	0.000**	0.000**
10y,1y,3y	0.000**	0.007**	0.000**	0.000**	0.000**
10y,1y,5y	0.000**	0.000**	0.000**	0.000**	0.000**
10y,1y,7y	0.000**	0.000**	0.000**	0.000**	0.000**
10y,3y,5y	0.000**	0.000**	0.000**	0.000**	0.000**
10y,3y,7y	0.000**	0.000**	0.000**	0.000**	0.000**
10y,5y,7y	0.000**	0.000**	0.000**	0.000**	0.000**
3m,1y,3y	0.080	0.000**	0.000**	0.000**	0.000**
3m,1y,5y	0.053	0.000**	0.000**	0.000**	0.000**
3m,1y,7y	0.037*	0.000**	0.000**	0.000**	0.000**
3m,3y,5y	0.000**	0.000**	0.000**	0.000**	0.000**
3m,3y,7y	0.000**	0.000**	0.000**	0.000**	0.001**
3m,5y,7y	0.000**	0.000**	0.000**	0.000**	0.000**
1y,3y,5y	0.000**	0.000**	0.000**	0.000**	0.000**
1y,3y,7y	0.000**	0.000**	0.000**	0.000**	0.003**
1y,5y,7y	0.000**	0.000**	0.000**	0.000**	0.000**
3y,5y,7y	0.000**	0.000**	0.000**	0.000**	0.000**

Table 5: In-sample tests of null hypothesis that contemporaneous values for Y_{2t} do not help predict other elements of Y_{2t} once Y_{1t} is included in the regression for alternative specifications of Y_{1t}. Table entries report p-values, with * denoting rejection at the 5% level, and ** rejection at the 1% level. Individual columns report predictability for individual elements of Y_{2t}. Regressions estimated 1983:M1-2002:M7.

Specification of Y_{1t}	Out-of-sample improvement in MSE				
	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}
6m,2y,10y	36%**	32%**	64%**	68%**	44%**
6m,2y,3m	98%**	-267%	89%**	98%**	98%**
6m,2y,1y	98%**	19%**	93%**	98%**	99%**
6m,2y,3y	92%**	22%**	8%**	80%**	91%**
6m,2y,5y	69%**	28%**	25%**	6%*	63%**
6m,2y,7y	29%**	27%**	32%**	48%**	51%**
6m,10y,3m	98%**	-17%**	99%**	94%**	79%**
6m,10y,1y	96%**	1%*	98%**	92%**	72%**
6m,10y,3y	47%**	45%**	46%**	28%**	29%**
6m,10y,5y	88%**	49%**	36%**	84%**	16%**
6m,10y,7y	96%**	56%**	43%**	97%**	82%**
6m,3m,1y	98%**	99%**	99%**	99%**	99%**
6m,3m,3y	71%**	97%**	-515%	94%**	97%**
6m,3m,5y	95%**	88%**	-155%**	95%**	76%**
6m,3m,7y	97%**	64%**	-73%**	97%**	81%**
6m,1y,3y	74%**	97%**	11%**	95%**	97%**
6m,1y,5y	93%**	86%**	4%**	93%**	75%**
6m,1y,7y	95%**	56%**	1%*	96%**	77%**
6m,3y,5y	8%**	63%**	37%**	33%**	63%**
6m,3y,7y	30%**	22%**	34%**	43%**	26%**
6m,5y,7y	91%**	46%**	26%**	-1%**	90%**
2y,10y,3m	79%**	73%**	75%**	74%**	54%**
2y,10y,1y	69%**	76%**	54%**	60%**	35%**
2y,10y,3y	94%**	87%**	76%**	-17%*	18%**
2y,10y,5y	96%**	90%**	85%**	43%**	17%**
2y,10y,7y	97%**	92%**	87%**	73%**	60%**
2y,3m,1y	16%**	98%**	92%**	98%**	99%**
2y,3m,3y	67%**	92%**	57%**	81%**	91%**
2y,3m,5y	75%**	71%**	67%**	29%**	62%**
2y,3m,7y	75%**	36%**	69%**	59%**	51%**
2y,1y,3y	56%**	91%**	71%**	82%**	92%**
2y,1y,5y	66%**	65%**	74%**	-6%	65%**
2y,1y,7y	71%**	17%**	75%**	41%**	50%**
2y,3y,5y	94%**	69%**	86%**	79%**	65%**
2y,3y,7y	94%**	30%**	85%**	79%**	12%**
2y,5y,7y	96%**	35%**	89%**	87%**	50%**
10y,3m,1y	13%**	96%**	98%**	92%**	74%**
10y,3m,3y	85%**	69%**	83%**	39%**	39%**
10y,3m,5y	88%**	93%**	86%**	87%**	20%**
10y,3m,7y	92%**	98%**	91%**	98%**	86%**
10y,1y,3y	76%**	17%**	80%**	15%**	21%**
10y,1y,5y	70%**	85%**	78%**	85%**	13%**
10y,1y,7y	71%**	93%**	78%**	95%**	78%**
10y,3y,5y	98%**	76%**	93%**	93%**	18%**
10y,3y,7y	98%**	80%**	94%**	94%**	29%**
10y,5y,7y	99%**	96%**	95%**	97%**	89%**
3m,1y,3y	-3%	73%**	97%**	95%**	97%**
3m,1y,5y	4%**	93%**	86%**	93%**	73%**
3m,1y,7y	4%**	95%**	59%**	96%**	76%**
3m,3y,5y	81%**	46%**	63%**	77%**	63%**
3m,3y,7y	80%**	55%**	25%**	78%**	25%**
3m,5y,7y	80%**	92%**	39%**	76%**	90%**
1y,3y,5y	71%**	-3%	64%**	76%**	65%**
1y,3y,7y	77%**	12%**	16%**	78%**	25%**
1y,5y,7y	63%**	91%**	49%**	68%**	91%**
3y,5y,7y	98%**	74%**	20%**	93%**	94%**

Table 6: Out-of-sample tests of null hypothesis that contemporaneous values for Y_{2t} do not help predict other elements of Y_{2t} once Y_{1t} is included in the regression for alternative specifications of Y_{1t}. Table entries report percent improvement in MSE for equation that includes $Y_{2t}^{(n)}$ over equation that does not for $Y_{2t}^{(n)}$ the elements of Y_{2t} other than that on the left-hand side of the regression. Asterisk (*) denotes Clark-West statistic leads to rejection of the null hypothesis of no improvement in the forecast at the 5% level, while ** denotes rejection at 1% level. Based on recursive regressions generating out-of-sample forecasts for 2002:M8-2007:M7.

Number of	Predictability tests				
principal components	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}
$m = 3$	0.000**	0.000**	0.000**	0.000**	0.000**
$m = 4$		0.000**	0.000**	0.000**	0.000**
$m = 5$			0.000**	0.000**	0.000**
$m = 6$				0.000**	0.000**
$m = 7$					0.000**

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, column j entry reports p-value for predicting the jth principal component z_{jt} when the contemporaneous values of the first m principal components are included. Regressions estimated 1983:M1-2002:M7.

Number of	Predictability tests				
principal components	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}
$m = 3$	75%**	6%**	44%**	65%**	39%**
$m = 4$		11%**	44%**	66%**	33%**
$m = 5$			41%**	71%**	34%**
$m = 6$				68%**	36%**
$m = 7$					36%**

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, column j entry reports the percent improvement in MSE for predicting the jth principal component z_{jt} when the contemporaneous values of the first m principal components are included. Asterisk (*) indicates that the Clark-West statistic leads to rejection of the null hypothesis of no improvement in the forecast at the 5% level, while ** denotes rejection at 1% level. Based on recursive regressions generating out-of-sample forecasts for 2002:M8-2007:M7.

lag length	interaction	likelihood ratio test	AIC	BIC
$k = 12$	dependent	–	**-3094**	-2038
$k = 12$	independent	$\chi^2(6) = 15.9$ $(p{=}0.0141)^*$	-3090	-2054
$k = 1$	dependent	$\chi^2(220) = 442.3265$ $(p{=}0.0000)^{**}$	-3090	-2783
$k = 1$	independent	$\chi^2(226) = 463.6942$ $(p{=}0.0000)^{**}$	-3080	**-2794**

Table 9: In-sample comparison of macro-finance models with different independence and lag length assumptions. First column reports likelihood ratio tests (p-value in parentheses) for testing indicated row against the first row. AIC = Akaike Information Criterion, and BIC = Schwarz Criterion, with bold indicating the preferred specification by that criterion. Regressions estimated 1983:M1-2002:M7.

lag length	interaction	6m	2y	10y	3m	1y	3y	5y	7y
$k = 12$	dependent	0.025	0.077	0.092	0.038	0.035	0.095	0.104	0.099
$k = 12$	independent	0.026	0.077	0.092	0.040	0.035	0.094	0.104	0.099
$k = 1$	dependent	**0.013**	**0.054**	**0.078**	**0.027**	0.024	**0.073**	**0.088**	**0.088**
$k = 1$	independent	0.015	0.059	0.082	0.028	**0.023**	0.079	0.093	0.092

Table 10: Post-sample comparison of macro finance models with different independence and lag length assumptions. Table entry is out-of-sample MSE for one-month-ahead forecast of the indicated yield on the basis of the indicated specification, with bold indicating the best out-of-sample performance for that variable. Based on recursive regressions generating out-of-sample forecasts for 2002:M8-2007:M7.