Identification and Estimation of Gaussian Affine Term Structure Models*

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

Jing (Cynthia) Wu‡
Department of Economics
University of California, San Diego

June 18, 2010
Revised: December 11, 2010

Abstract

This paper develops new results for both identification and estimation of Gaussian affine term structure models. In terms of identification, we establish that three popular canonical representations are each, for different reasons, unidentified. We also demonstrate that a failure of local identification can complicate numerical search for the maximum-likelihood estimate when one uses conventional estimation methods. A separate contribution of the paper is the proposal of minimum-chi-square estimation as an alternative to maximum-likelihood estimation. We show that, although it is asymptotically equivalent or sometimes even identical to MLE, it can be much easier to compute. In some cases, MCSE allows the researcher to recognize with certainty whether a given estimate represents a global maximum of the likelihood function and makes feasible the computation of small-sample standard errors.

*We are grateful to Bryan Brown, Michael Bauer and Ken Singleton for comments on an earlier draft.
†jhamilton@ucsd.edu
‡jingwu@ucsd.edu

1 Introduction.

The class of Gaussian affine term structure models[1] developed by Vasicek (1977), Duffie and Kan (1996), Dai and Singleton (2002), and Duffee (2002) has become the basic workhorse in macroeconomics and finance for purposes of using a no-arbitrage framework for studying the relations between yields on assets of different maturities. Its appeal comes from its simple characterization of how risk gets priced by the market which, under the assumption of no arbitrage, generates predictions for the price of any asset. The approach has been used to measure the role of risk premia in interest rates (Duffee, 2002; Cochrane and Piazzesi, 2009), study how macroeconomic developments and monetary policy affect the term structure of interest rates (Ang and Piazzesi, 2003; Beechey and Wright, 2009; Bauer, 2009), characterize the monetary policy rule (Ang, Dong and Piazzesi, 2007; Rudebusch and Wu, 2008; Bekaert, Cho and Moreno, 2010), determine why long-term yields remained remarkably low in 2004 and 2005 (Kim and Wright, 2005; Rudebusch, Swanson and Wu, 2006), infer market expectations of inflation from the spread between nominal and inflation-indexed Treasury yields (Christensen, Lopez and Rudebusch, 2010), evaluate the effectiveness of the extraordinary central bank interventions during the financial crisis (Christensen, Lopez and Rudebusch, 2009; Smith, 2010), and study the potential for monetary policy to affect interest rates when the short rate is at the zero lower bound Hamilton and Wu (2010a).

But buried in the footnotes of this literature and in the practical experience of those who have used these models are tremendous numerical challenges in estimating the necessary

[1] By Gaussian affine term structure models we refer to specifications in which the discrete-time joint distribution of yields and factors is multivariate Normal with constant conditional variances. We do not in this paper consider the broader class of non-Gaussian processes.

parameters from the data due to highly non-linear and badly-behaved likelihood surfaces. For example, Kim (2008) observed:

> Flexibly specified no-arbitrage models tend to entail much estimation difficulty due to a large number of parameters to be estimated and due to the nonlinear relationship between the parameters and yields that necessitates a nonlinear optimization.

Ang and Piazzesi (2003) similarly reported:

> difficulties associated with estimating a model with many factors using maximum likelihood when yields are highly persistent....We need to find good starting values to achieve convergence in this highly non-linear system....[T]he likelihood surface is very flat in λ_0 which determines the mean of long yields....

This paper proposes a solution to these and other problems with affine term structure models based on what we will refer to as their reduced-form representation. For a popular class of Gaussian affine term structure models– namely, those for which the model is claimed to price exactly a subset of N_ℓ linear combinations of observed yields, where N_ℓ is the number of unobserved pricing factors– this reduced form is a restricted vector autoregression in the observed set of yields and macroeconomic variables. More generally, the reduced form is a restricted state-space representation for the set of observed variables. We explore two implications of this fact that seem to have been ignored in the large preceding literature on such models.

The first is that the parameters of these reduced-form representations contain all the observable implications of any Gaussian affine term structure model for the sample of observed data. Hence, as noted by Fisher (1966) and Rothenberg (1971), we can use the reduced-form representation to characterize the identifiability of any parameters that might be of interest. If more than one value for the parameter vector of interest is associated with the same reduced-form parameter vector, then the model is unidentified at that point and there is no way to use the observed data to distinguish between the alternative possibilities. We show using this approach that three popular parameterizations of affine term structure models– namely, the preferred representations proposed by Dai and Singleton (2000), Ang and Piazzesi (2003) and Pericoli and Taboga (2008)– are in fact unidentified. While the lack of identification of the Dai and Singleton (2000) representation has previously been established by Collin-Dufresne, Goldstein and Jones (2008) and Aït-Sahalia and Kimmel (2010) using other methods, the results for the Ang and Piazzesi (2003) and Pericoli and Taboga (2008) approaches are new. We further demonstrate that it is common for numerical search methods to end up in regions of the parameter space that are locally unidentified, and show why this failure of identification arises. These issues of identification are one factor that contributes to the numerical difficulties for conventional methods noted above.

A second and completely separate contribution of the paper is the observation that it is possible for the parameters of interest to be inferred directly from estimates of the reduced-form parameters themselves. This is a very useful result because the latter are often simple OLS coefficients. Although translating from reduced-form parameters into structural parameters involves a mix of analytical and numerical calculations, the numerical component is

far simpler than that associated with the usual approach of trying to find the maximum of the likelihood surface directly as a function of the structural parameters. In the case of a just-identified structure, the numerical component of our proposed method has an additional big advantage over the traditional approach, in that the researcher knows with certainty whether the equations have been solved, and therefore knows with certainty whether one has found the global maximum of the likelihood surface with respect to the structural parameters or simply a local maximum. In the conventional approach, one instead has to search over hundreds of different starting values, and even then has no guarantee that the global maximum has been found. In the case where the model imposes overidentifying restrictions on the reduced form, one can still estimate structural parameters as functions of the unrestricted reduced-form estimates by the method of minimum-chi-square estimation described by Rothenberg (1973, pp. 24-25).[2] This minimizes a quadratic form in the difference between the reduced-form parameters implied by a given structural model and the reduced-form parameters as estimated without restrictions directly from the data, with weights coming from the information matrix. Among other illustrations of the computational benefits of this approach, we establish the feasibility of calculating small-sample standard errors and confidence intervals for this class of models and demonstrate that the parameter estimates reported by Ang and Piazzesi (2003) in fact correspond to a local maximum of the likelihood surface and are not the global MLE.

There have been several other recent efforts to address many of these problems. Christensen, Diebold and Rudebusch (forthcoming) develop a no-arbitrage representation of a dynamic Nelson-Siegel model of interest rates that gives a convenient representation of level,

[2]Minimum-chi-square estimation has also been used in other settings by Chamberlain (1982) and Newey (1987).

slope and curvature factors and offers significant improvements in empirical tractability and predictive performance over earlier affine term structure specifications. Joslin, Singleton and Zhu (forthcoming) propose a canonical representation for affine term structure models that greatly improves convergence of maximum likelihood estimation. Collin-Dufresne, Goldstein and Jones (2008) propose a representation in terms of the derivatives of the term structure at maturity zero, arguing for the benefits of using these observable magnitudes rather than unobserved latent variables to represent the state vector of an ATSM. Each of these papers proposes canonical representations that are identified, and the Christensen, Diebold and Rudebusch (forthcoming) and Joslin, Singleton and Zhu (forthcoming) parameterizations lead to better behaved likelihood functions than do the parameterizations explored in detail in our paper.

The chief difference between our proposed solution and those of these other researchers is that they focus on how the ATSM should be represented, whereas we examine how the parameters of the ATSM are to be estimated. Thus for example Christensen, Diebold and Rudebusch (forthcoming) require the researcher to impose certain restrictions on the ATSM, whereas Joslin, Singleton and Zhu (forthcoming) cannot incorporate most auxiliary restrictions on the P dynamics. It is far from clear how any of these three approaches could have been used to estimate a model of the form investigated by Ang and Piazzesi (2003). By contrast, our minimum-chi-square algorithm can be used for any representation, including those proposed by Christensen, Diebold and Rudebusch (forthcoming) and Joslin, Singleton and Zhu (forthcoming), and can simplify the numerical burden regardless of the representation chosen. Indeed, some of the numerical advantages of Joslin, Singleton and Zhu (forthcom-

ing) come from the fact that a subset of their parameterization is identical to a subset of our reduced-form representation, and their approach, like ours, takes advantage of the fact that the full-information MLE for this subset can be obtained by OLS for a popular class of models. However, Joslin, Singleton and Zhu (forthcoming) estimate the remaining parameters by conventional MLE rather than using the full set of reduced-form estimates as in our approach. As Joslin, Singleton and Zhu (forthcoming) note, their representation becomes unidentified in the presence of a unit root. When applied to highly persistent data, we illustrate that their MLE algorithm can encounter similar problems to those of other representations, which can be avoided with our approach to parameter estimation.

Our estimation strategy is related to that of Bekaert, Cho and Moreno (2010), who estimate structural parameters to match the moments of the reduced-form representation using the generalized method of moments (GMM). Whereas our approach uses OLS estimation to simplify greatly the numerical estimation problem, their approach requires all parameters to be estimated together from scratch in a single GMM problem that retains all of the numerical challenges associated with MLE.

The rest of the paper is organized as follows. Section 2 describes the class of Gaussian affine term structure models and three popular examples, and briefly uses one of the specifications to illustrate the numerical difficulties that can be encountered with the traditional approach. Section 3 investigates the mapping from structural to reduced-form parameters. We establish that the canonical forms of all three examples are unidentified and explore how this contributes to some of the problems for conventional numerical search algorithms. In Section 4 we use the mapping to propose approaches to parameter estimation that are much better behaved.

Section 5 concludes.

2 Gaussian Affine Term Structure Models.

2.1 Basic framework

Consider an $(M \times 1)$ vector of variables F_t whose dynamics are characterized by a Gaussian vector autoregression:

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

with $u_t \sim$ i.i.d. $N(0, I_M)$. This specification implies that $F_{t+1}|F_t, F_{t-1}, ..., F_1 \sim N(\mu_t, \Sigma\Sigma')$ for

$$\mu_t = c + \rho F_t. \tag{2}$$

Let r_t denote the risk-free one-period interest rate. If the vector F_t includes all the variables that could matter to investors, then the price of a pure discount asset at date t should be a function $P_t(F_t)$ of the current state vector. Moreover, if investors were risk neutral, the price they'd be willing to pay would satisfy

$$\begin{aligned} P_t(F_t) &= \exp(-r_t) E_t \left[P_{t+1}(F_{t+1}) \right] \\ &= \exp(-r_t) \int_{\mathbb{R}^M} P_{t+1}(F_{t+1}) \phi(F_{t+1}; \mu_t, \Sigma\Sigma') \, dF_{t+1} \end{aligned} \tag{3}$$

for $\phi(y;\mu,\Omega)$ the M-dimensional $N(\mu,\Omega)$ density evaluated at the point y:

$$\phi(y;\mu,\Omega) = \frac{1}{(2\pi)^{M/2}|\Omega|^{1/2}} \exp\left[-\frac{(y-\mu)'\Omega^{-1}(y-\mu)}{2}\right]. \tag{4}$$

More generally, with risk-averse investors we would replace (3) with

$$\begin{aligned} P_t(F_t) &= E_t\left[P_{t+1}(F_{t+1})M_{t,t+1}\right] \\ &= \int_{\mathbb{R}^M} P_{t+1}(F_{t+1})\left[M_{t,t+1}\phi(F_{t+1};\mu_t,\Sigma\Sigma')\right]\, dF_{t+1} \end{aligned} \tag{5}$$

for $M_{t,t+1}$ the pricing kernel. In many macro models, the pricing kernel would be

$$M_{t,t+1} = \frac{\beta U'(C_{t+1})}{U'(C_t)(1+\pi_{t+1})}$$

for β the personal discount rate, $U'(C)$ the marginal utility of consumption, and π_{t+1} the inflation rate between t and $t+1$.

Affine term structure models are derived from the particular kernel

$$M_{t,t+1} = \exp\left[-r_t - (1/2)\lambda_t'\lambda_t - \lambda_t' u_{t+1}\right] \tag{6}$$

for λ_t an $(M \times 1)$ vector that characterizes investor attitudes toward risk, with $\lambda_t = 0$ in the case of risk neutrality. Elementary multiplication of (4) by (6) reveals that for this case

$$M_{t,t+1}\phi(F_{t+1};\mu_t,\Sigma\Sigma') = \exp(-r_t)\phi(F_{t+1};\mu_t^Q,\Sigma\Sigma') \tag{7}$$

for

$$\mu_t^Q = \mu_t - \Sigma\lambda_t. \tag{8}$$

Substituting (7) into (5) and comparing with (3), we see that for this specification of the pricing kernel, risk-averse investors value any asset the same as risk-neutral investors would if the latter thought that the conditional mean of F_{t+1} was μ_t^Q rather than μ_t. A positive value for the first element of λ_t, for example, implies that an asset that delivers the quantity $F_{1,t+1}$ dollars in period $t+1$ would have a value at time t that is less than the value that would be assigned by a risk-neutral investor, and the size of this difference is bigger when the $(1,1)$ element of Σ is bigger. An asset yielding $F_{i,t+1}$ dollars has a market value that is reduced by $\Sigma_{i1}\lambda_{1t}$ relative to a risk-neutral valuation, through the covariance between factors i and 1. The term λ_{1t} might then be described as the market price of factor 1 risk.

The affine term structure models further postulate that this market price of risk is itself an affine function of F_t,

$$\lambda_t = \lambda + \Lambda F_t \tag{9}$$

for λ an $(M \times 1)$ vector and Λ an $(M \times M)$ matrix. Substituting (9) and (2) into (8), we see that

$$\mu_t^Q = c^Q + \rho^Q F_t$$

for

$$c^Q = c - \Sigma\lambda \tag{10}$$

$$\rho^Q = \rho - \Sigma\Lambda. \tag{11}$$

In other words, risk-averse investors value assets the same way as a risk-neutral investor would if that risk-neutral investor believed that the factors are characterized by a Q-measure VAR given by

$$F_{t+1} = c^Q + \rho^Q F_t + \Sigma u_{t+1}^Q \tag{12}$$

with u_{t+1}^Q a vector of independent standard Normal variables under the Q measure.

Suppose that the risk-free 1-period yield is also an affine function of the factors

$$r_t = \delta_0 + \delta_1' F_t. \tag{13}$$

Then, as demonstrated for example in Appendix A of Ang and Piazzesi (2003), under the above assumptions the yield on a risk-free n-period pure-discount bond can be calculated as

$$y_t^n = a_n + b_n' F_t \tag{14}$$

where

$$b_n = \frac{1}{n}\left[I_M + \rho^{Q\prime} + \cdots + \left(\rho^{Q\prime}\right)^{n-1}\right]\delta_1 \tag{15}$$

$$\begin{aligned} a_n = {} & \delta_0 + \left(b_1' + 2b_2' + \cdots + (n-1)\, b_{n-1}'\right) c^Q/n \\ & - \left(b_1'\Sigma\Sigma' b_1 + 2^2 b_2'\Sigma\Sigma' b_2 + \cdots + (n-1)^2\, b_{n-1}'\Sigma\Sigma' b_{n-1}\right)/2n. \end{aligned} \tag{16}$$

If we knew F_t and the values of c^Q and ρ^Q along with δ_0, δ_1, and Σ, we could use (14), (15), and (16) to predict the yield for any maturity n.

There are thus three sets of parameters that go into an affine term structure model: (a)

the parameters c, ρ, and Σ that characterize the objective dynamics of the factors in equation (1) (sometimes called the P parameters); (b) the parameters λ and Λ in equation (9) that characterize the price of risk; and (c) the Q parameters c^Q and ρ^Q (along with the same Σ as appeared in the P parameter set) that figure in (12). If we knew any two of these sets of parameters, we could calculate the third[3] using (10) and (11). We will refer to a representation in terms of (a) and (b) as a λ representation, and a representation in terms of (a) and (c) as a Q representation.

Suppose we want to describe yields on a set of N_d different maturities. If N_d is greater than N_ℓ, where N_ℓ is the number of unobserved pricing factors, then (14) would imply that it should be possible to predict the value of one of the y_{nt} as an exact linear function of the others. Although in practice we can predict one yield extremely accurately given the others, the empirical fit is never exact. One common approach to estimation, employed for example by Ang and Piazzesi (2003) and Chen and Scott (1993), is to suppose that (14) holds exactly for N_ℓ linear combinations of observed yields, and that the remaining $N_e = N_d - N_\ell$ linear combinations differ from the predicted value by a small measurement error. Let Y_t^1 denote the $(N_\ell \times 1)$ vector consisting of those linear combinations of yields that are treated as priced without error and Y_t^2 the remaining $(N_e \times 1)$ linear combinations. The measurement

[3]We will discuss examples below in which Σ is singular for which the demonstration of this equivalence is a bit more involved, with the truth of the assertion coming from the fact that for such cases certain elements of λ and Λ are defined to be zero.

specification is then

$$\begin{bmatrix} Y_t^1 \\ {\scriptstyle (N_\ell \times 1)} \\ Y_t^2 \\ {\scriptstyle (N_e \times 1)} \end{bmatrix} = \begin{bmatrix} A_1 \\ {\scriptstyle (N_\ell \times 1)} \\ A_2 \\ {\scriptstyle (N_e \times 1)} \end{bmatrix} + \begin{bmatrix} B_1 \\ {\scriptstyle (N_\ell \times M)} \\ B_2 \\ {\scriptstyle (N_e \times M)} \end{bmatrix} F_t + \begin{bmatrix} 0 \\ {\scriptstyle (N_\ell \times N_e)} \\ \Sigma_e \\ {\scriptstyle (N_e \times N_e)} \end{bmatrix} \underset{(N_e \times 1)}{u_t^e} \tag{17}$$

where Σ_e is typically taken to be diagonal. Here A_i and B_i are calculated by stacking (16) and (15), respectively, for the appropriate n, while Σ_e determines the variance of the measurement error with $u_t^e \sim N(0, I_{N_e})$. We will discuss many of the issues associated with identification and estimation of affine term structure models in terms of three examples.

2.2 Example 1: Latent factor model.

In this specification, the factors F_t governing yields are treated as if observable only through their implications for the yields themselves; examples in the continuous-time literature include Dai and Singleton (2000), Duffee (2002), and Kim and Orphanides (2005). Typically in this case, the number of factors N_ℓ and the number of yields observed without error are both taken to be 3, with the 3 factors interpreted as the level, slope, and curvature of the term structure. The 3 linear combinations Y_t^1 regarded as observed without error can be constructed from the first 3 principal components of the set of yields. Alternatively, they could be constructed directly from logical measures of level, slope, and curvature. Yet another option is simply to choose 3 representative yields as the elements of Y_t^1. Which linear combinations are claimed to be priced without error can make a difference for certain testable implications of the model, an issue that is explored in a separate paper by Hamilton and Wu (2010b) which

addresses empirical testing of the overidentifying restrictions of affine term structure models. For purposes of discussing identification and estimation, however, the choice of which yields go into Y_t^1 is immaterial, and notation is kept simplest by following Ang and Piazzesi (2003) and Pericoli and Taboga (2008) in just using 3 representative yields. In our numerical example, these are taken to be the $n = 1$-, 12-, and 60-month maturities, with data on 36-month yields included separately in Y_t^2. Thus for this illustrative latent-factor specification, equation (17) takes the form

$$\begin{bmatrix} y_t^1 \\ y_t^{12} \\ y_t^{60} \\ y_t^{36} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_{12} \\ a_{60} \\ a_{36} \end{bmatrix} + \begin{bmatrix} b_1' \\ b_{12}' \\ b_{60}' \\ b_{36}' \end{bmatrix} F_t + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \Sigma_e \end{bmatrix} u_t^e \tag{18}$$

where a_n and b_n are calculated from equations (16) and (15), respectively.

We will use for our illustration a Q representation for this system. Dai and Singleton (2000) proposed the normalization conditions $\Sigma = I_{N_\ell}$, $\delta_1 \geq 0$, $c = 0$ and ρ lower triangular. Singleton (2006) used parallel constraints on the Q parameters ($\Sigma = I_{N_\ell}$, $\delta_1 \geq 0$, $c^Q = 0, \rho^Q$ lower triangular). Our illustration will use $\Sigma = I_{N_\ell}$, $\delta_1 \geq 0$, $c = 0$ and ρ^Q lower triangular. For the $N_\ell = 3$, $N_e = 1$ case displayed in equation (18), there are then 23 unknown parameters: 3 in c^Q, 6 in ρ^Q, 9 in ρ, 1 in δ_0, 3 in δ_1, and 1 in Σ_e, which we collect in the (23×1) vector θ. The log likelihood is

$$\mathcal{L}(\theta; Y) = \sum_{t=1}^{T} \{-\log[|\det(J)|] + \log \phi(F_t; c + \rho F_{t-1}, I_{N_\ell}) + \log \phi(u_t^e; 0, I_{N_e})\} \tag{19}$$

for $\phi(.)$ the multivariate Normal density in equation (4) and $\det(J)$ the determinant of the

Jacobian, with

$$J = \begin{bmatrix} \underset{(N_\ell \times N_\ell)}{B_1} & \underset{(N_\ell \times N_e)}{0} \\ \underset{(N_e \times N_\ell)}{B_2} & \underset{(N_e \times N_e)}{\Sigma_e} \end{bmatrix}$$

$$F_t = B_1^{-1}(Y_t^1 - A_1)$$

$$u_t^e = \Sigma_e^{-1}\left\{Y_t^2 - A_2 - B_2 B_1^{-1}(Y_t^1 - A_1)\right\}.$$

The Chen-Scott procedure is to maximize (19) with respect to θ by numerical search.

As a simple example to illustrate the difficulties with this traditional estimation and some of the advantages of the procedure that we will be recommending to replace it, we simulated a sample of 1000 observations using parameters specified in the first block of Table 1 below. These parameters were chosen to match the actual observed behavior of the four yields used here. On this sample we tried to choose θ so as to maximize (19) using the fminunc algorithm in MATLAB.[4] Since numerical search can be sensitive to different scaling of parameters, we tried to scale parameters in a way consistent with a researcher's prior expectation that risk prices were small, multiplying c^Q by 10 and δ_1 and Σ_e by 1000 so that a unit step for each of these parameters would be similar to a unit step for the others.[5] We used 100 different starting values for this search, using a range of values for ρ^Q and starting the other parameters at good guesses. Specifically, to obtain a given starting value we would generate the 3 diagonal

[4]MATLAB numerical optimizers have been used by Cochrane and Piazzesi (2009), Aït-Sahalia and Kimmel (2010), and Joslin, Singleton and Zhu (forthcoming), among others. Duffee (2009) found that numerical search problems can be reduced using alternative algorithms. Our purpose here is to illustrate the difficulties that can arise in estimation. We will demonstrate that these identical MATLAB algorithms have no trouble with the alternative formulation that we will propose below.

[5]To give the algorithm the best chance to converge, for each starting value we allowed the search to continue for up to 10,000 function evaluations, then restarted the search at that terminal value to allow an additional 10,000 function evaluations, and so on, for 10 repetitions with each starting value.

elements of ρ^Q from $U[0.5, 1]$ distributions, set off-diagonal elements to zero, and set the initial guess for ρ equal to this value for ρ^Q. We set the starting value for each element of δ_1 and Σ_e to 1.e-4, $\delta_0 = 0.0046$ (the average short rate), and $c^Q = 0$.

	True values			Global maximum			Local 53		
c^Q	0.0407	0.0135	0.5477	0.0416	0.0085	0.5316	-0.5562	0.0204	0.0527
ρ^Q	0.9991	0	0	0.9985	0	0	0.9986	0	0
	0.0101	0.9317	0	0.0116	0.9328	0	0.0113	0.9316	0
	0.0289	0.2548	0.7062	0.0219	0.2500	0.7202	0.0203	0.2438	0.7352
ρ	0.9812	0.0069	0.0607	0.9696	0.0141	0.0671	0.9794	0.0063	0.0840
	-0.0010	0.8615	0.1049	-0.0027	0.8533	0.1175	-0.0028	0.8380	0.1267
	0.0164	0.1856	0.6867	0.0085	0.1985	0.6993	0.0333	0.1923	0.7202
δ_0	0.0046			0.0046			0.1344		
δ_1	1.729E-4	1.803E-4	4.441E-4	1.71E-4	1.71E-4	4.45E-4	1.72E-4	1.59E-4	4.54E-4
Σ_e	9.149E-5			9.105E-5			9.110E-5		
eig(ρ)	0.9879	0.9341	0.6074	0.9734	0.9448	0.6040	1.000	0.9306	0.6070
LLF				28110.4			28096.5		

Table 1: Parameter values used for simulation and estimates associated with (1) the global maximum and (2) a representative point of local convergence.

In only 1 of these 100 experiments did the numerical search converge to the values that we will establish below are indeed the true global MLE. These estimates, reported in the second block of Table 1, in fact correspond very nicely to the true values from which this sample was simulated. However, in 81 of the other experiments, the procedure satisfied the convergence criterion (usually coming from a sufficiently tiny change between iterations) at a large range of alternative points other than the global maximum. The third block of Table 1 displays one of these. All such points are characterized by an eigenvalue of ρ being equal or very close to unity; we will explain why this happens in the following section. For the other 18 starting values, the search algorithm was unable to make any progress from the initial starting values. Although very simple, this exercise helps convey some sense of the numerical

problems researchers have encountered fitting more complicated models such as we describe in our next two examples.

2.3 Example 2: Macro finance model with single lag (MF1).

It is of considerable interest to include observable macroeconomic variables among the factors that may affect interest rates, as for example in Ang and Piazzesi (2003), Ang, Dong and Piazzesi (2007), Rudebusch and Wu (2008), Ang, Piazzesi and Wei (2006), and Hordahl, Tristani and Vestin (2006). Our next two illustrative examples come from this class. We first consider the unrestricted first-order macro factor model studied by Pericoli and Taboga (2008). This model uses $N_m = 2$ observable macro factors, consisting of measures of the inflation rate and the output gap, which are collected in an $(N_m \times 1)$ vector f_t^m. These two observable macroeconomic factors are allowed to influence yield dynamics in addition to the traditional $N_\ell = 3$ latent[6] factors f_t^ℓ,

$$\underset{(N_f \times 1)}{F_t} = \begin{bmatrix} \underset{(N_m \times 1)}{f_t^m} \\ \\ \underset{(N_\ell \times 1)}{f_t^\ell} \end{bmatrix},$$

[6] Pericoli and Taboga evaluated a number of alternative specifications including different choices for the number of latent factors N_ℓ, number of lags on the macro variables, and dependence between the latent and macro factors. They refer to the specification we discuss in the text as the $M(3, 0, U)$ specification, which is the one that their tests suggest best fits the data.

for $N_f = N_m + N_\ell$. The P dynamics (1), Q dynamics (12), and short-rate equation (13) can for this example be written in partitioned form as

$$\begin{aligned} \underset{(N_m \times 1)}{f_t^m} &= c_m + \rho_{mm} f_{t-1}^m + \rho_{m\ell} f_{t-1}^\ell + \Sigma_{mm} u_t^m \\ \underset{(N_\ell \times 1)}{f_t^\ell} &= c_\ell + \rho_{\ell m} f_{t-1}^m + \rho_{\ell\ell} f_{t-1}^\ell + \Sigma_{\ell m} u_t^m + \Sigma_{\ell\ell} u_t^\ell \end{aligned} \tag{20}$$

$$\begin{aligned} \underset{(N_m \times 1)}{f_t^m} &= c_m^Q + \rho_{mm}^Q f_{t-1}^m + \rho_{m\ell}^Q f_{t-1}^\ell + \Sigma_{mm} u_t^{Qm} \\ \underset{(N_\ell \times 1)}{f_t^\ell} &= c_\ell^Q + \rho_{\ell m}^Q f_{t-1}^m + \rho_{\ell\ell}^Q f_{t-1}^\ell + \Sigma_{\ell m} u_t^{Q_m} + \Sigma_{\ell\ell} u_t^{Q\ell} \end{aligned} \tag{21}$$

$$r_t = \delta_0 + \delta_{1m}' f_t^m + \delta_{1\ell}' f_t^\ell. \tag{22}$$

Pericoli and Taboga proposed the normalization conditions[7] that Σ_{mm} is lower triangular, $\Sigma_{\ell m} = 0$, $\Sigma_{\ell\ell} = I_{N_\ell}$, $\delta_{1\ell} \geq 0$, and $c_\ell^Q = 0$.

Our empirical illustration of this approach will use t corresponding to quarterly data and will take the 1-, 5-, and 10-year bonds to be priced without error ($Y_t^1 = (y_t^4, y_t^{20}, y_t^{40})'$) and the 2-, 3-, and 7-year bonds to be priced with error ($Y_t^2 = (y_t^8, y_t^{12}, y_t^{28})'$). Details of how the log likelihood is calculated for this example are described in Appendix A.

[7] Pericoli and Taboga imposed $f_0^\ell = 0$ as an alternative to the traditional $c_\ell = 0$ or $c_\ell^Q = 0$, though we will follow the rest of the literature here in using a more standard normalization.

2.4 Example 3: Macro finance model with 12 lags (MF12).

A first-order VAR is not sufficient to capture the observed dynamics of output and inflation. For example, Ang and Piazzesi (2003) suggested that the best fit is obtained using a monthly VAR(12) in the observable macro variables and a VAR(1) for the latent factors:[8]

$$\begin{aligned} f_t^m &= \rho_1 f_{t-1}^m + \rho_2 f_{t-2}^m + \cdots + \rho_{12} f_{t-12}^m + \Sigma_{mm} u_t^m \\ f_t^\ell &= c_\ell + \rho_{\ell\ell} f_{t-1}^\ell + \Sigma_{\ell\ell} u_t^\ell. \end{aligned}$$

Our empirical example follows Ang and Piazzesi in proxying the 2 elements of f_t^m with the first principal components of a set of output and and a set of inflation measures, respectively, which factors have mean zero by construction. Ang and Piazzesi treated the macro dynamics as independent of those for the unobserved latent factors, so that terms such as $\rho_{\ell m}$ and $\rho_{m\ell}$ in the preceding example are set to zero.

Ang and Piazzesi (2003) further proposed the following identifying restrictions: Σ_{mm} is lower triangular, $\Sigma_{\ell\ell} = I_{N_\ell}$, $c_\ell = 0$, $\rho_{\ell\ell}$ is lower triangular, and the diagonal elements of $\rho_{\ell\ell}$ are in descending order. Further restrictions and details of the model and its likelihood function are provided in Appendix B. In the specification we replicate, Ang and Piazzesi postulated that the short rate depends only on the current values of the macro factors:

$$r_t = \delta_0 + \delta_{1m}' f_t^m + \delta_{1\ell}' f_t^\ell.$$

[8] Ang and Piazzesi refer to this as their Macro Model.

They further noted that since f_t^ℓ is independent of f_t^m under their assumptions, the values of δ_0 and δ_{1m} in the short-rate equation can be obtained by OLS estimation of

$$r_t = \delta_0 + \delta_{1m}^{'} f_t^m + v_t. \tag{23}$$

To further reduce the dimensionality of the estimation, Ang and Piazzesi (2003) proposed some further restrictions on this set-up that we will discuss in more detail in Section 4.4.

3 Identification.

The log likelihood function for each of the models discussed– and indeed, for any Gaussian affine term structure model in which exactly N_ℓ linear combinations of yields are assumed to be priced without error– takes the form of a restricted vector autoregression. The mapping from the affine-pricing parameters to the VAR parameters allows us to evaluate the identifiability of a given structure. If two different values for the structural parameters imply the identical reduced-form parameters, there is no way to use observable data to choose between the two. We now explore the implications of this fact for each of the three classes of models described in the previous section.

3.1 Example 1: Latent factor model.

Premultiplying (1) by B_1 (and recalling the normalization $c = 0$ and $\Sigma = I_{N_\ell}$) results in

$$B_1 F_t = B_1 \rho B_1^{-1} B_1 F_{t-1} + B_1 u_t.$$

Adding A_1 to both sides and substituting $Y_t^1 = A_1 + B_1 F_t$ establishes

$$Y_t^1 = A_1^* + \phi_{11}^* Y_{t-1}^1 + u_{1t}^* \tag{24}$$

$$A_1^* = A_1 - B_1 \rho B_1^{-1} A_1 \tag{25}$$

$$\phi_{11}^* = B_1 \rho B_1^{-1}. \tag{26}$$

Likewise the second block of (17) implies

$$Y_t^2 = A_2^* + \phi_{21}^* Y_t^1 + u_{2t}^* \tag{27}$$

$$A_2^* = A_2 - B_2 B_1^{-1} A_1 \tag{28}$$

$$\phi_{21}^* = B_2 B_1^{-1} \tag{29}$$

$$\begin{bmatrix} u_{1t}^* \\ u_{2t}^* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega_1^* & 0 \\ 0 & \Omega_2^* \end{bmatrix} \right) \tag{30}$$

$$\Omega_1^* = B_1 B_1' \tag{31}$$

$$\Omega_2^* = \Sigma_e \Sigma_e'. \tag{32}$$

Equations (24) and (27) will be recognized as a restricted Gaussian VAR for Y_t, in which a single lag of Y_{t-1}^1 appears in the equation for Y_t^1 and in which, after conditioning on the contemporaneous value of Y_t^1, no lagged terms appear in the equation for Y_t^2. Note that when we refer to the reduced-form for this system, we will incorporate those exclusion restrictions

along with the restriction that Ω_2^* is diagonal.

Table 2 summarizes the mapping between the VAR parameters and the affine term structure parameters implied by equations (24)-(32).[9] The number of VAR parameters minus the number of structural parameters is equal to $(N_e - 1)(N_\ell + 1)$. Thus the structure is just-identified by a simple parameter count when $N_e = 1$ and overidentified when $N_e > 1$. Notwithstanding, the structural parameters can nevertheless be unidentified despite the apparent conclusion from a simple parameter count.

VAR parameter	No. of elements	Σ_e N_e	ρ^Q $N_\ell(N_\ell+1)/2$	δ_1 N_ℓ	ρ N_ℓ^2	c^Q N_ℓ	δ_0 1
Ω_2^*	N_e	✓					
ϕ_{21}^*	$N_\ell N_e$		✓				
Ω_1^*	$N_\ell(N_\ell+1)/2$		✓	✓			
ϕ_{11}^*	N_ℓ^2		✓	✓	✓		
A_2^*	N_e		✓	✓		✓	✓
A_1^*	N_ℓ		✓	✓	✓	✓	✓

Table 2: Mapping between structural and reduced-form parameters for the latent factor model.

Consider first what happens at a point where one of the eigenvalues of ρ is unity, that is, when the P-measure factor dynamics exhibit a unit root.[10] This means that one of the eigenvalues of $B_1\rho B_1^{-1}$ is also unity ($B_1\rho B_1^{-1}x = x$ for some nonzero x) requiring that $(I_{N_\ell} - B_1\rho B_1^{-1})x = 0$, so the matrix $I_{N_\ell} - B_1\rho B_1^{-1}$ is noninvertible. In this case, even if we knew the true value of A_1^*, we could never find the value of A_1 from equation (25). If $\hat{A}_1$ is proposed as a fit for a given sample, then $\hat{A}_1 + kx$ produces the identical fit for any k. Note moreover from (16) that A_1 and A_2 are the only way to find out about c^Q and δ_0; if we don't

[9] The value of δ_1 turns out not to appear in the product $\phi_{21}^* = B_2 B_1^{-1}$.

[10] Note we have followed Ang and Piazzesi (2003) and Joslin, Singleton and Zhu (forthcoming), among others, in basing estimates on the likelihood function conditional on the first observation. By contrast, Chen and Scott (1993) and Duffee (2002) include the unconditional likelihood of the first observation as a device for imposing stationarity.

know the 4 values in A_1 and A_2, we can never infer the 4 values of c^Q and δ_0. This failure of local identification accounts for the numerous failed searches described in Section 2.2. When the search steps in a region in which ρ has a near unit root, the likelihood surface becomes extremely flat in one direction (and exactly flat at the unit root), causing the numerical search to become bogged down. Because the true process is quite persistent, it is extremely common for a numerical search to explore this region of the surface and become stuck.[11]

If instead we used the normalization $c^Q = 0$ in place of the condition $c = 0$ just analyzed, a similar phenomenon occurs in which a unit root in ρ^Q results in a failure of local identification of δ_0.

Even when all eigenvalues of ρ are less than unity, there is another respect in which the latent factor model discussed here is unidentified.[12] Let H denote any $(N_\ell \times N_\ell)$ matrix such that $H'H = I_{N_\ell}$. It is apparent from equations (24)-(32) that if we replace B_j by $B_j H'$ and ρ by $H\rho H'$, there would be no change in the implied value for the sample likelihood. The question then is whether the conditions imposed on the underlying model rule out such a transformation. From equation (16), such a transformation requires replacing c^Q with Hc^Q, and from (15) we need now to use $H\delta_1$ and $H\rho^Q H'$. Since our specification imposed no restrictions on ρ or c^Q, the question is whether the proposed lower triangular structure for ρ^Q and nonnegativity of δ_1 rules out such a transformation. The following proposition establishes that it does not.

[11] This point has also been made by Aït-Sahalia and Kimmel (2010).

[12] This has also been recognized by Ang and Piazzesi (2003), Collin-Dufresne, Goldstein and Jones (2008) and Aït-Sahalia and Kimmel (2010).

Proposition 1. *Consider any* (2×2) *lower triangular matrix:*

$$\rho^Q = \begin{bmatrix} \rho_{11}^Q & 0 \\ \rho_{21}^Q & \rho_{22}^Q \end{bmatrix}.$$

Then for almost all (2×1) *positive vectors* δ_1*, there exists a unique orthogonal matrix* H *other than the identity matrix such that* $H\rho^Q H'$ *is also lower triangular and* $H\delta_1 > 0$*. Moreover,* $H\rho^Q H'$ *takes one of the following forms:*

$$\begin{bmatrix} \rho_{22}^Q & 0 \\ \rho_{21}^Q & \rho_{11}^Q \end{bmatrix} \text{ or } \begin{bmatrix} \rho_{22}^Q & 0 \\ -\rho_{21}^Q & \rho_{11}^Q \end{bmatrix}.$$

For ρ^Q *an* $(N_\ell \times N_\ell)$ *lower triangular matrix, there are* $N_\ell!$ *different lower triangular representations, characterized by alternative orderings of the principal diagonal elements.*

There thus exist 6 different parameter configurations that would achieve the same maximum for the likelihood function for the latent example explored in Section 2.2. The experiment did not uncover them because the other difficulties with maximization were sufficiently severe that for the 100 different starting values used, only one of these 6 configurations was reached. Dai and Singleton (2000) and Singleton (2006) originally proposed lower triangularity of ρ or ρ^Q and nonnegativity of δ_1 as sufficient identifying conditions. Our proposition establishes that one needs a further condition such as $\rho_{11}^Q \geq \rho_{22}^Q \geq \rho_{33}^Q$ to have a globally identified structure.

Nevertheless, this multiplicity of global optima is a far less serious problem than the failure of local identification arising from a unit root. The reason is that any of the alternative

configurations obtained through these H transformations by construction has the identical implications for bond pricing. By contrast, the inferences one would draw from Local 53 in Table 1 are fundamentally flawed and introduce substantial practical difficulties for using this class of models.

There is another identification issue, which has separately been recognized by Joslin, Singleton and Zhu (forthcoming) using a very different approach from ours: not all matrices ρ^Q can be transformed into lower triangular form. For example, for $N_\ell = 2$, if ρ^Q is written as lower triangular, then ρ^Q_{22} would have to be one of its eigenvalues. However, it is possible for an unrestricted real-valued matrix ρ^Q to have complex eigenvalues, in which case there is no way to transform it as $\Upsilon = H\rho^Q H'$ for Υ a real-valued lower triangular matrix. We propose in the following proposition an alternative normalization for the case $N_\ell = 2$ that, unlike the usual lower-triangular form, is completely unrestrictive.

Proposition 2. *Consider ρ^Q any (2×2) real-valued matrix:*

$$\rho^Q = \begin{bmatrix} \rho^Q_{11} & \rho^Q_{12} \\ \rho^Q_{21} & \rho^Q_{22} \end{bmatrix}.$$

For almost all $\delta_1 \in \mathbb{R}^{2+}$, there exist exactly two transformations of the form $\Upsilon = H\rho^Q H'$ such that Υ is real, $H'H = I_2$, $H\delta_1 > 0$, and the two elements on the principal diagonal of Υ are the same. Moreover, one of these transformations is simply the transpose of the other:

$$\Upsilon_1 = \begin{bmatrix} a & b \\ c & a \end{bmatrix} \qquad \Upsilon_2 = \begin{bmatrix} a & c \\ b & a \end{bmatrix}.$$

Hence one approach for the $N_\ell = 2$ case would be to choose the 3 parameters a, b, and c so as to maximize the likelihood with

$$\rho^Q = \begin{bmatrix} a & b \\ c & a \end{bmatrix}$$

subject to the normalization $b \leq c$. This has the advantage over the traditional lower-triangular formulation in that the latter imposes additional restrictions on the dynamics (namely, lower-triangular ρ^Q rules out the possibility of complex roots) whereas the Υ formulation does not.

Unfortunately, it is less clear how to generalize this to larger dimensions. If ρ^Q has complex eigenvalues, these always appear as complex conjugates. Thus if one knew for the case $N_\ell = 3$ that ρ^Q contained complex eigenvalues, a natural normalization would be

$$\rho^Q = \begin{bmatrix} \rho^Q_{11} & 0 & 0 \\ \rho^Q_{21} & a & \rho^Q_{23} \\ \rho^Q_{31} & \rho^Q_{32} & a \end{bmatrix} \tag{33}$$

with $\rho^Q_{23} \leq \rho^Q_{32}$ The value of a is then uniquely pinned down by the real part of the complex eigenvalues. However, if the eigenvalues are all real, this is a more awkward form than the usual

$$\rho^Q = \begin{bmatrix} \rho^Q_{11} & 0 & 0 \\ \rho^Q_{21} & \rho^Q_{22} & 0 \\ \rho^Q_{31} & \rho^Q_{32} & \rho^Q_{33} \end{bmatrix} \tag{34}$$

with $\rho_{11}^Q \geq \rho_{22}^Q \geq \rho_{33}^Q$. The estimation approach that we propose below will instantly reveal whether or not the lower triangular form (34) is imposing a restriction relative to the full-information maximum likelihood unrestricted values. If (34) is determined not to impose a restriction, one can feel confident in using the conventional parameterization, whereas if it does turn out to be inconsistent with the estimated unrestricted dynamics, the researcher should instead parameterize dynamics using (33).

3.2 Example 2: Macro finance model with single lag.

We next examine the MF1 specification of Pericoli and Taboga (2008). Calculations similar to those for the latent factor model show the reduced form to be

$$\underset{(N_m\times 1)}{f_t^m} = \underset{(N_m\times 1)}{A_m^*} + \underset{(N_m\times N_m)}{\phi_{mm}^*} f_{t-1}^m + \underset{(N_m\times N_\ell)}{\phi_{m1}^*} Y_{t-1}^1 + u_{mt}^* \tag{35}$$

$$\underset{(N_\ell\times 1)}{Y_t^1} = \underset{(N_\ell\times 1)}{A_1^*} + \underset{(N_\ell\times N_m)}{\phi_{1m}^*} f_{t-1}^m + \underset{(N_\ell\times N_\ell)}{\phi_{11}^*} Y_{t-1}^1 + \underset{(N_\ell\times N_m)}{\psi_{1m}^*} f_t^m + u_{1t}^* \tag{36}$$

$$\underset{(N_e\times 1)}{Y_t^2} = \underset{(N_e\times 1)}{A_2^*} + \underset{(N_e\times N_m)}{\phi_{2m}^*} f_t^m + \underset{(N_e\times N_\ell)}{\phi_{21}^*} Y_t^1 + u_{2t}^*. \tag{37}$$

Once again it is convenient to include the contemporaneous value of f_t^m in the equation for Y_t^1 and include contemporaneous values of both f_t^m and Y_t^1 in the equation for Y_t^2 in order to orthogonalize the reduced-form residuals u_{jt}^*; the benefits of this representation will be seen in the next section. The mapping between structural and reduced-form parameters is given

by the following equations and summarized in Table 3 with $N_f = N_m + N_\ell$:

$$
\begin{aligned}
A_m^* &= c_m - \rho_{m\ell} B_{1\ell}^{-1} A_1 && (38)\\
\phi_{mm}^* &= \rho_{mm} - \rho_{m\ell} B_{1\ell}^{-1} B_{1m} && (39)\\
\phi_{m1}^* &= \rho_{m\ell} B_{1\ell}^{-1} && (40)\\
A_1^* &= A_1 + B_{1\ell} c_\ell - B_{1\ell} \rho_{\ell\ell} B_{1\ell}^{-1} A_1 && (41)\\
\phi_{1m}^* &= B_{1\ell} \rho_{\ell m} - B_{1\ell} \rho_{\ell\ell} B_{1\ell}^{-1} B_{1m} && (42)\\
\phi_{11}^* &= B_{1\ell} \rho_{\ell\ell} B_{1\ell}^{-1} && (43)\\
\psi_{1m}^* &= B_{1m} && (44)\\
A_2^* &= A_2 - B_{2\ell} B_{1\ell}^{-1} A_1 && (45)\\
\phi_{2m}^* &= B_{2m} - B_{2\ell} B_{1\ell}^{-1} B_{1m} && (46)\\
\phi_{21}^* &= B_{2\ell} B_{1\ell}^{-1} && (47)
\end{aligned}
$$

$$
\text{Var}\begin{bmatrix} u_{mt}^* \\ u_{1t}^* \\ u_{2t}^* \end{bmatrix} = \begin{bmatrix} \Omega_m^* & 0 & 0 \\ 0 & \Omega_1^* & 0 \\ 0 & 0 & \Omega_2^* \end{bmatrix} = \begin{bmatrix} \Sigma_{mm}\Sigma_{mm}' & 0 & 0 \\ 0 & B_{1\ell} B_{1\ell}' & 0 \\ 0 & 0 & \Sigma_e \Sigma_e' \end{bmatrix} \qquad (48)
$$

with Ω_2^* diagonal and B_1 and B_2 partitioned as described in Appendix A.

Once again inspection of the above equations reveals that the structure is unidentified. One can see this immediately for the case $N_\ell = 3$, $N_m = 2$, $N_e = 3$ simply by counting parameters– there are 69 unknown structural parameters and only 66 reduced-form parameters from which they are supposed to be inferred. The problem arises in particular from the fact

VAR parameter	No. of elements	Σ_e	Σ_{mm}	ρ^Q	δ_1	$\rho_{m\ell}$	ρ_{mm}	$\rho_{\ell\ell}$	$\rho_{\ell m}$	δ_0	c^Q	c_m	c_ℓ
		N_e	$N_m(N_m+1)/2$	N_f^2	N_f	N_mN_ℓ	N_m^2	N_ℓ^2	$N_\ell N_m$	1	N_m	N_m	N_ℓ
Ω_2^*	N_e	✓											
Ω_m^*	$N_m(N_m+1)/2$		✓										
ψ_{1m}^*	$N_\ell N_m$			✓	✓								
ϕ_{2m}^*	$N_e N_m$			✓	✓								
ϕ_{21}^*	$N_e N_\ell$			✓	✓								
Ω_1^*	$N_\ell(N_\ell+1)/2$			✓	✓								
ϕ_{m1}^*	$N_m N_\ell$			✓	✓	✓							
ϕ_{mm}^*	N_m^2			✓	✓	✓	✓						
ϕ_{11}^*	N_ℓ^2			✓	✓			✓					
ϕ_{1m}^*	$N_\ell N_m$			✓	✓			✓	✓				
A_2^*	N_e		✓	✓	✓					✓	✓		
A_m^*	N_m		✓	✓	✓	✓				✓	✓	✓	
A_1^*	N_ℓ		✓	✓	✓			✓		✓	✓		✓

Table 3: Mapping between structural and reduced-form parameters for the MF1 model.

that, for the example we have been discussing, the observable implications of the 30 structural parameters in ρ^Q and δ_1 are completely captured by the 27 values of ψ_{1m}^*, ϕ_{2m}^*, ϕ_{21}^*, and Ω_1^*. More fundamentally, the lack of identification would remain with this structure no matter how large the value of N_e. One can see this by verifying that the following transformation is perfectly allowed under the stated normalization but would not change the value of any reduced-form parameter: $B_{1\ell} \to B_{1\ell}H'$, $c_\ell \to Hc_\ell$, $\rho_{m\ell} \to \rho_{m\ell}H'$, $\rho_{\ell\ell} \to H\rho_{\ell\ell}H'$, $\rho_{\ell m} \to H\rho_{\ell m}$, and $B_{2\ell} \to B_{2\ell}H'$, where H could be any $(N_\ell \times N_\ell)$ orthogonal matrix.

There is also a separate identification problem arising from the fact that only maturities for which n is an even number are included in the observation set. This means that only even powers of ρ^Q appear in (15) and (16), which allows observationally equivalent sign transformations through H as well.

3.3 Example 3: Macro finance model with 12 lags.

Last we consider the MF12 example, for which the reduced form is

$$\underset{(2\times1)}{f_t^m} = \underset{(2\times24)}{\phi_{mm}^*} F_{t-1}^m + u_{mt}^* \tag{49}$$

$$\underset{(3\times1)}{Y_t^1} = A_1^* + \underset{(3\times24)}{\phi_{1m}^*} F_{t-1}^m + \underset{(3\times3)}{\phi_{11}^*} Y_{t-1}^1 + \underset{(3\times2)}{\psi_{1m}^*} f_t^m + u_{1t}^* \tag{50}$$

$$\underset{(2\times1)}{Y_t^2} = A_2^* + \underset{(2\times24)}{\phi_{2m}^*} F_t^m + \underset{(2\times3)}{\phi_{21}^*} Y_t^1 + u_{2t}^* \tag{51}$$

$$\begin{aligned}
\phi_{mm}^* &= \begin{bmatrix} \rho_1 & \rho_2 & \cdots & \rho_{12} \end{bmatrix} \\
A_1^* &= A_1 - B_{1\ell}\rho_{\ell\ell}B_{1\ell}^{-1}A_1 \\
\underset{(3\times24)}{\phi_{1m}^*} &= \begin{bmatrix} \underset{(3\times22)}{B_{1m}^{(1)}} & \underset{(3\times2)}{0} \end{bmatrix} - \underset{(3\times3)}{B_{1\ell}\rho_{\ell\ell}B_{1\ell}^{-1}} \begin{bmatrix} \underset{(3\times2)}{B_{1m}^{(0)}} & \underset{(3\times22)}{B_{1m}^{(1)}} \end{bmatrix} \\
\phi_{11}^* &= B_{1\ell}\rho_{\ell\ell}B_{1\ell}^{-1} \\
\psi_{1m}^* &= B_{1m}^{(0)} \\
A_2^* &= A_2 - B_{2\ell}B_{1\ell}^{-1}A_1 \\
\phi_{2m}^* &= B_{2m} - B_{2\ell}B_{1\ell}^{-1}B_{1m} \\
\phi_{21}^* &= B_{2\ell}B_{1\ell}^{-1}
\end{aligned}$$

$$\text{Var}\left(\begin{bmatrix} u_{mt}^* \\ u_{1t}^* \\ u_{2t}^* \end{bmatrix}\right) = \begin{bmatrix} \Omega_m^* & 0 & 0 \\ 0 & \Omega_1^* & 0 \\ 0 & 0 & \Omega_2^* \end{bmatrix} = \begin{bmatrix} \Sigma_{mm}\Sigma_{mm}' & 0 & 0 \\ 0 & B_{1\ell}B_{1\ell}' & 0 \\ 0 & 0 & \Sigma_e\Sigma_e' \end{bmatrix}$$

with Ω_2^* again diagonal and details on the partitioning of B_1 and B_2 in Appendix B. Table 4 summarizes the mapping between reduced-form and structural parameters. Note that the

only reduced-form parameters relevant for inference about the 6 elements of δ_0 and λ are the 5 values for A_1^* and A_2^*, establishing that these structural parameters are in fact unidentified. One might have thought that perhaps δ_0 could be inferred separately from the OLS regression (23), freeing up the parameters A_1^* and A_2^* for estimation solely of λ. However, this is not the case, since the short-term interest rate is the same dependent variable in both regression (23) and in the first OLS regression from which A_1^* is inferred. Another way to see this is to note that at most what one can expect to uncover from the 5 values of A_1^* and A_2^* are the 5 values of A_1 and A_2. The first element of A_1 is exactly equal to δ_0, so even if δ_0 were known a priori, the most that one could infer from A_1 and A_2 is 4 other parameters. Hence A_1 and A_2 would not be sufficient to uncover the 5 unknowns in λ even if δ_0 were known with certainty.

VAR	No. of	Σ_e	Σ_{mm}	$\rho_{1,\ldots,12}$	Λ_{mm}	δ_{1m}	$\rho_{\ell\ell}$	$\Lambda_{\ell\ell}$	$\delta_{1\ell}$	δ_0	λ
parameter	elements	2	3	48	4	2	6	9	3	1	5
Ω_2^*	2	✓									
Ω_m^*	3		✓								
ϕ_{mm}^*	48			✓							
ψ_{1m}^*	6			✓	✓	✓					
ϕ_{21}^*	6						✓	✓	✓		
Ω_1^*	6						✓	✓	✓		
ϕ_{11}^*	9						✓	✓	✓		
ϕ_{2m}^*	48			✓	✓	✓	✓	✓	✓		
ϕ_{1m}^*	72			✓	✓	✓	✓	✓	✓		
A_2^*	2		✓	✓	✓	✓	✓	✓	✓	✓	✓
A_1^*	3		✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 4: Mapping between structural and reduced-form parameters for the MF12 model.

Ang and Piazzesi's (2003) Macro Model with its proposed identifying restrictions thus turns out to be unidentified at all points of the parameter space. In their empirical analysis, Ang and Piazzesi imposed an additional set of restrictions that were intended to improve estimation efficiency, though as we have just seen some of these are necessary for identification. We

discuss these further in Section 4.4 below.

4 Estimation.

The reduced-form parameters are trivially obtained via OLS. Hence a very attractive alternative to numerical maximization of the log likelihood function directly with respect to the structural parameters θ is to let OLS do the work of maximizing the likelihood with respect to the reduced-form parameters, and then translate these into their implications for θ. We demonstrate in this section how this can be done.

4.1 Minimum-chi-square estimation.

Let π denote the vector consisting of reduced-form parameters (VAR coefficients and nonredundant elements of the variance matrices), $\mathcal{L}(\pi; Y)$ denote the log likelihood for the entire sample, and $\hat{\pi} = \arg\max \mathcal{L}(\pi; Y)$ denote the full-information-maximum-likelihood estimate. If $\hat{R}$ is a consistent estimate of the information matrix,

$$R = -T^{-1} E \left[\frac{\partial^2 \mathcal{L}(\pi; Y)}{\partial \pi \ \partial \pi'} \right]$$

then we could test the hypothesis that $\pi = g(\theta)$ for θ a known vector of parameters by calculating the usual Wald statistic

$$T \left[\hat{\pi} - g(\theta) \right]' \hat{R} \left[\hat{\pi} - g(\theta) \right] \tag{52}$$

which would have an asymptotic $\chi^2(q)$ distribution under the null hypothesis where q is the dimension of π. Rothenberg (1973, p. 24) noted that one could also use (52) as a basis for estimation by choosing as an estimate $\hat{\theta}$ the value that minimizes this chi-square statistic.

Following Rothenberg (1973, pp. 24-25), we can obtain asymptotic standard errors by considering the linear approximation $g(\theta) \simeq \gamma + \Gamma\theta$ for $\Gamma = \partial g(\theta)/\partial\theta'|_{\theta=\theta_0}$ and $\gamma = g(\theta_0) - \Gamma\theta_0$ where $\hat{\pi} \xrightarrow{p} \pi_0$ and we assume there exists a value of θ_0 for which the true model satisfies $g(\theta_0) = \pi_0$. Define the linearized minimum-chi-square estimator $\hat{\theta}^*$ as the solution to

$$\min_{\theta} \; T\left[\hat{\pi} - \gamma - \Gamma\theta\right]' R \left[\hat{\pi} - \gamma - \Gamma\theta\right],$$

that is, $\hat{\theta}^*$ satisfies $\Gamma' R(\hat{\pi} - \gamma - \Gamma\hat{\theta}^*) = 0$ or $\hat{\theta}^* = (\Gamma' R\Gamma)^{-1}\Gamma' R(\hat{\pi} - \gamma)$. Since $\sqrt{T}(\hat{\pi} - \pi_0) \xrightarrow{L} N(0, R^{-1})$, it follows that $\sqrt{T}(\hat{\theta}^* - \theta_0) \xrightarrow{L} N\left(0, [\Gamma' R\Gamma]^{-1}\right)$. Hence our proposal is to approximate the variance of $\hat{\theta}$ with $T^{-1}(\hat{\Gamma}'\hat{R}\hat{\Gamma})^{-1}$ for $\hat{\Gamma} = \partial g(\theta)/\partial\theta'|_{\theta=\hat{\theta}}$.

We show in Appendix E that this is in fact identical to the usual asymptotic variance for the MLE as obtained from second derivatives of the log likelihood function directly with respect to θ. In other words, the MCSE and MLE are asymptotically equivalent, and the MCSE inherits all the asymptotic optimality properties of the MLE. If in a particular sample the MCSE and MLE differ, there is no basis for claiming that one has better properties than the other.

In the case of a just-identified model, the minimum value attainable for (52) is zero, in which case one can without loss of generality simply minimize

$$\left[\hat{\pi} - g(\theta)\right]' \left[\hat{\pi} - g(\theta)\right]. \tag{53}$$

Note that in this case, if the optimized value for this objective is zero, then $\hat{\theta}$ is numerically identical to the value that achieves the global maximum of the likelihood written as a function of θ. Although $\hat{\theta}_{MCSE}$ in this case is identical to $\hat{\theta}_{MLE}$, arriving at the estimate by the minimum-chi-square algorithm has two big advantages over the traditional brute-force maximization of the likelihood function. First, one knows instantly whether $\hat{\theta}$ corresponds to a global maximum of the original likelihood surface simply by checking whether a zero value is achieved for (53). By contrast, under the traditional approach, one has to try hundreds of starting values to be persuaded that a global maximum has been found, and even then cannot be sure. A second advantage is that minimization of (52) or (53) is far simpler computationally than brute-force maximization of the original likelihood function.

In addition, the greater computational ease makes calculation of small-sample confidence intervals feasible. The models considered here imply a reduced form that can be written in companion form as

$$Y_t = k + \Phi Y_{t-1} + \Sigma_Y u_t$$

for Y_t the $(N \times 1)$ vector of observed variables (yields, macro variables, and possible lags of macro variables) and $u_t \sim N(0, I_N)$, where the parameters k, Φ, and Σ_Y are known functions of π. We can then obtain bootstrap confidence intervals for θ as follows. For artificial sample j, we will generate a sequence $\{u_t^{(j)}\}_{t=1}^{T}$ of $N(0, I_N)$ variables for T the original sample size, and then recursively generate $Y_t^{(j)} = k(\hat{\pi}) + \Phi(\hat{\pi})Y_{t-1}^{(j)} + \Sigma_Y(\hat{\pi})u_t^{(j)}$ for $t = 1, 2, ..., T$, starting from $Y_0^{(j)} = Y_0$, the initial value from the original sample, and using the identical parameter values k, Φ, and Σ_Y (as implied by the original $\hat{\pi}$) for each sample j. On sample j we find the FIML estimate $\hat{\pi}^{(j)}$ on that artificial sample and then calculate

$\hat{\theta}^{(j)} = \arg\min_{\theta} T\left[\hat{\pi}^{(j)} - g(\theta)\right]' \hat{R}^{(j)} \left[\hat{\pi}^{(j)} - g(\theta)\right]$. We generate a sequence $j = 1, 2, ..., J$ of such samples, from which we could calculate 95% small-sample confidence intervals for each element of θ. The small-sample standard errors for parameter i reported in the following section were calculated from $\sqrt{J^{-1}\sum_{j=1}^{J}(\hat{\theta}_{i,MCSE}^{(j)} - \hat{\theta}_i)^2}$ where $\hat{\theta}_i$ is the MCSE estimate for the original sample (whose original FIML $\hat{\pi}$ was used to generate each artificial sample j) and $\hat{\theta}_{i,MCSE}^{(j)}$ is the minimum-chi-square estimate for artificial sample j.

We now illustrate these methods and their advantages in detail using the examples of affine term structure models discussed above.

4.2 Example 1: Latent factor model.

In the case of $N_e = 1$, the latent factor model is just-identified, making application of minimum-chi-square estimation particularly attractive. The reduced-form parameter vector here is

$$\pi = \left(\left\{ \text{vec}\left(\begin{bmatrix} A_1^* & \phi_{11}^* \end{bmatrix}' \right) \right\}', [\text{vech}(\Omega_1^*)]', \left\{ \text{vec}\left(\begin{bmatrix} A_2^* & \phi_{21}^* \end{bmatrix}' \right) \right\}', [\text{diag}(\Omega_2^*)]' \right)'$$

where $\text{vec}(X)$ stacks the columns of the matrix X into a vector. If X is square, $\text{vech}(X)$ does the same using only the elements on or below the principal diagonal, and $\text{diag}(X)$ constructs a vector from the diagonal elements of X. Because u_{1t}^* and u_{2t}^* are independent, full-information-maximum-likelihood (FIML) estimation of π is obtained by treating the Y_1 and Y_2 blocks separately. Since each equation of (24) has the same explanatory variables, FIML for the ith row of $[A_1^*, \phi_{11}^*]$ is obtained by OLS regression of Y_{it}^1 on a constant and Y_{t-1}^1, with $\hat{\Omega}_1^*$ the

matrix of average outer products of those OLS residuals:

$$\hat{\Omega}_1^* = T^{-1}\sum_{t=1}^{T}(Y_t^1 - \hat{A}_1^* - \hat{\phi}_{11}^* Y_{t-1}^1)(Y_t^1 - \hat{A}_1^* - \hat{\phi}_{11}^* Y_{t-1}^1)'.$$

FIML estimates of the remaining elements of π are likewise obtained from OLS regressions of Y_{it}^2 on a constant and Y_t^1.

The specific mapping in Table 2 suggests that we can use the following multi-step algorithm to minimize (53) for the latent factor model with $N_\ell = 3$ and $N_e = 1$.

Step 1. The estimate of Σ_e is obtained analytically from the square root of $\hat{\Omega}_2^*$.

Step 2. The estimates of the 9 unknowns in ρ^Q and δ_1 are found by numerically solving the 9 equations in (29) and (31)

$$[B_2(\hat{\rho}^Q, \hat{\delta}_1)][B_1(\hat{\rho}^Q, \hat{\delta}_1)]' = \hat{\phi}_{21}^* \hat{\Omega}_1^*$$

$$[B_1(\hat{\rho}^Q, \hat{\delta}_1)][B_1(\hat{\rho}^Q, \hat{\delta}_1)]' = \hat{\Omega}_1^*.$$

Specifically, we do this by letting[13] $\hat{\pi}_2 = (\left[\text{vec}(\hat{\phi}_{21}^* \hat{\Omega}_1^*)\right]', [\text{vech}(\hat{\Omega}_1^*)]')'$ and $g_2(\rho^Q, \delta_1) = ([\text{vec}(B_2 B_1')]',$ $[\text{vech}(B_1 B_1')]')'$ and finding $\hat{\rho}^Q$ and $\hat{\delta}_1$ by numerical minimization of $[\hat{\pi}_2 - g_2(\rho^Q, \delta_1)]'[\hat{\pi}_2 - g_2(\rho^Q, \delta_1)]$.

[13] To assist with scaling for numerical robustness, we multiplied each equation in step 2 by $1200 \times 1.\text{e+}7$ and those in step 4 below by 1.e+8. If we were minimizing (52) directly one would automatically achieve optimal scaling by using $\hat{R}$ in place of a constant k times the identity matrix as here. However, our formulation takes advantage of the fact that the elements of $\hat{\pi}$ can be rearranged in order to avoid inversion of B_1 inside the numerical optimization, in which case $\hat{R}$ is no longer the optimal weighting matrix. The minimization was implemented using the fsolve command in MATLAB. We also multiplied δ_1 by 1000 to improve numerical robustness.

Step 3. The estimate of ρ can then be obtained analytically from (26):

$$\hat{\rho} = \hat{B}_1^{-1} \hat{\phi}_{11}^* \hat{B}_1 \tag{54}$$

where $\hat{B}_1$ is known from Step 2.

Step 4. Numerically solve the 4 unknowns in δ_0 and c^Q from the 4 equations in $\hat{A}_1^*$ and $\hat{A}_2^*$ using (25) and (28):

$$\left(I_3 - \hat{B}_1 \hat{\rho} \hat{B}_1^{-1}\right) A_1(\delta_0, c^Q, \hat{\rho}^Q, \hat{\delta}_1) = \hat{A}_1^*$$

$$A_2(\delta_0, c^Q, \hat{\rho}^Q, \hat{\delta}_1) - \hat{B}_2 \hat{B}_1^{-1} A_1(\delta_0, c^Q, \hat{\rho}^Q, \hat{\delta}_1) = \hat{A}_2^*.$$

Although Steps 2 and 4 involve numerical minimization, these are computationally far simpler problems than that associated with traditional brute-force maximization of the likelihood function with respect to the full vector θ. To illustrate this, we repeated the experiment described in Section 2.2 with the same 100 starting values. Whereas we saw in Section 2.2 that only one of these efforts found the global maximum under the traditional approach, with our method all 100 converge to the global MLE in one of the 6 configurations that are observationally equivalent for the original normalization. One of the reasons for the greater robustness is that the critical stumbling block for the traditional method– numerical search over ρ– is completely avoided since in our approach (54) is solved analytically. Another is that c^Q and uncertainties about its scale are completely eliminated from the core problem of estimation of ρ^Q and δ_1.

Joslin, Singleton and Zhu (forthcoming) have recently proposed a promising alternative parameterization of the pure latent affine models that shares some of the advantages of our approach. They parameterize the system such that A_1^* and ϕ_{11}^* in (24) are taken to be the direct objects of interest, and as in our approach, estimate these directly with OLS. But whereas our approach also uses the OLS estimates of A_2^* and ϕ_{21}^* in (27) to uncover the remaining affine-pricing parameters, their approach finds these by maximizing the joint likelihood function of Y_1 and Y_2. Although they report that the second step involves no numerical difficulties, our experience is that while it offers a significant improvement over the traditional method, it is still susceptible to some of the same problems. For example, we repeated the experiment described above with the same data set and same starting values for δ_0 and the 3 unknown diagonal elements in ρ^Q that appear in their parameterization as we used in the simulations described above, starting the search for Ω_1^* from the OLS estimates as they recommend. We found that the algorithm found the global maximum in 54 out of the 100 trials[14], but got stuck in regions with diagonal elements of ρ^Q equal to unity in the others, in a similar failure of local identification that we documented above can plague the traditional approach.

We applied our method directly to the Ang and Piazzesi interest rate data described in more detail in Section 4.4 below. Table 5 reports the resulting minimum-chi-square estimates (identical in this case to the full-information-maximum-likelihood estimates). The table also reports asymptotic standard errors in parentheses and small-sample standard errors in square brackets. The latter were calculated by applying our method to each of 1000 separate data

[14] To assist the numerical search, we multiplied Ω_1^* by 1000. Without this scaling, the searches only succeeded in finding the global maximum in 14 of the 100 trials.

sets, each generated from the vector autoregression estimated from the original data set. Note that the fact that we can verify with certainty that the global maximum has been found on each of these 1000 simulated data sets is part of what makes calculation of small-sample standard errors feasible and attractive. Finding the FIML estimate on 1000 data sets takes about 90 seconds on a PC. For this example, we find that the asymptotic standard errors provide an excellent approximation to the true small-sample values.

Although our original inference was conducted in terms of a Q representation, we report the implied λ representation values in the right-hand columns of Table 5, since that is the form in which parameter estimates are often reported for these models. Our suggestion is that the approach we illustrate here, of beginning with a completely unrestricted model to see which parameters appear to be most significant, has many advantages over the traditional approach[15] in which sundry restrictions are imposed at a very early stage, partly in order to assist with identification and estimation.

4.3 Example 2: Macro finance model with single lag.

We also applied this procedure to estimate parameters for our MF1 example using a slightly different quarterly data set from Pericoli and Taboga. We used constant-maturity Treasury yields as of the first day of the quarter, dividing the numbers as usually reported by 400 in order to convert to units of quarterly yield on which formulas such as (14) are based. We estimated inflation from the 12-month percentage change in the CPI and the output gap by applying the Hodrick-Prescott filter with $\lambda = 1600$ to 100 times the natural log of real GDP.

[15] See for example Duffee (2002) and Duarte (2004).

Estimated Q representation parameters				Implied λ representation parameters			
c^Q	0.0407	0.0135	0.5477	λ	-0.0407	-0.0135	-0.5477
	[0.0063]	[0.0399]	[0.1194]		[0.0063]	[0.0399]	[0.1194]
	(0.0062)	(0.0378)	(0.1073)				
ρ^Q	0.9991	0	0	Λ	-0.0178	0.0069	0.0607
	[0.0005]				[0.0109]	[0.0231]	[0.0303]
	(0.0004)						
	0.0101	0.9317	0		-0.0111	-0.0701	0.1049
	[0.0033]	[0.0050]			[0.0102]	[0.0323]	[0.0331]
	(0.0032)	(0.0046)					
	0.0289	0.2548	0.7062		-0.0125	-0.0693	-0.0195
	[0.0193]	[0.0206]	[0.0507]		[0.0090]	[0.0354]	[0.0449]
	(0.0185)	(0.0172)	(0.0439)				
ρ	0.9812	0.0069	0.0607				
	[0.0110]	[0.0231]	[0.0303]				
	(0.0067)	(0.0226)	(0.0294)				
	-0.0010	0.8615	0.1049				
	[0.0113]	[0.0343]	[0.0331]				
	(0.0094)	(0.0309)	(0.0318)				
	0.0164	0.1856	0.6867				
	[0.0187]	[0.0289]	[0.0353]				
	(0.0174)	(0.0277)	(0.0350)				
δ_0	0.0046						
	[0.0011]						
	(0.0011)						
δ_1	1.729E-4	1.803E-4	4.441E-4				
	[2.31E-5]	[3.80E-5]	[1.75E-5]				
	(2.28E-5)	(3.74E-5)	(1.62E-5)				
Σ_e	9.149E-5						
	[2.81E-6]						
	(2.70E-6)						

Table 5: FIML estimates with small-sample standard errors (in square brackets) and asymptotic standard errors (in parentheses) for latent factor model fit to Ang and Piazzesi (2003) data set.

Data run from 1960:Q1 to 2007:Q1 and were obtained from the FRED database of the Federal Reserve Bank of St. Louis.

If we impose 3 further restrictions on $\rho^Q_{\ell\ell}$ relative to the original formulation, the MF1 model presented above would be just-identified in terms of parameter count, for which we would logically again simply try to invert the reduced-form parameter estimates to obtain the FIML estimates of the structural parameters. Once again orthogonality of the residuals across the three blocks of (35) through (37) means FIML estimation can be done on each block separately, and within each block implemented by OLS equation by equation. Our estimation procedure on this system is then as follows.

Step 1. The f^m_t and Y^2_t variance parameters are obtained analytically from (48), that is, $\hat{\Sigma}_{mm}$ from the Cholesky factorization of $\hat{\Omega}^*_m$ and $\hat{\Sigma}_e$ from the square root of $\hat{\Omega}^*_2$.

Step 2. Using (44), (46), (48), and (47), choose the values of ρ^Q and δ_1 so as to solve the following equations numerically[16]:

$$B_{1m}(\rho^Q, \delta_1) = \hat{\psi}^*_{1m}$$

$$B_{2m}(\rho^Q, \delta_1) = \hat{\phi}^*_{2m} + \hat{\phi}^*_{21}\hat{\psi}^*_{1m}$$

$$\text{vech}\left\{\left[B_{1\ell}(\rho^Q, \delta_1)\right]\left[B_{1\ell}(\rho^Q, \delta_1)\right]'\right\} = \text{vech}\left(\hat{\Omega}^*_1\right)$$

$$\left[B_{2\ell}(\rho^Q, \delta_1)\right]\left[B_{1\ell}(\rho^Q, \delta_1)\right]' = \hat{\phi}^*_{21}\hat{\Omega}^*_1.$$

We initially tried to solve this system for $\rho^Q_{\ell\ell}$ of the lower-triangular form (34), but found no

[16] To improve accuracy of the numerical algorithm, we multiplied the last two equations by 400 and then the whole set of equations by 1.e+7. The parameter δ_1 was also scaled by 100.

solution exists, indicating that the FIML estimate of $\rho^Q_{\ell\ell}$ has complex roots. We accordingly reparameterized $\rho^Q_{\ell\ell}$ in the form (33), for which an exact solution was readily obtained.

Step 3. From these estimates one then analytically can calculate $\hat{\rho}_{m\ell}$, $\hat{\rho}_{mm}$, $\hat{\rho}_{\ell\ell}$, and $\hat{\rho}_{\ell m}$ from $\hat{\phi}^*_{m1}$, $\hat{\phi}^*_{mm}$, $\hat{\phi}^*_{11}$, and $\hat{\phi}^*_{1m}$, respectively.

Step 4. Since c_m and c_ℓ are unrestricted, the values of δ_0 and c^Q can be inferred solely from A^*_2 by numerical solution of (45):

$$A_2(\delta_0, c^Q, \hat{\rho}^Q, \hat{\delta}_1) - \hat{B}_{2\ell}\hat{B}^{-1}_{1\ell}\hat{A}_1(\delta_0, c^Q, \hat{\rho}^Q, \hat{\delta}_1) = \hat{A}^*_2.$$

Step 5. We then can calculate the remaining parameters analytically using (38) and (41):

$$\hat{c}_m = \hat{A}^*_m + \hat{\rho}_{m\ell}\hat{B}^{-1}_{1\ell}\hat{A}_1$$

$$\hat{c}_\ell = \hat{B}^{-1}_{1\ell}\left(\hat{A}^*_1 - \hat{A}_1 + \hat{B}_{1\ell}\hat{\rho}_{\ell\ell}\hat{B}^{-1}_{1\ell}\hat{A}_1\right).$$

Table 6 reports the FIML estimates obtained by the above algorithm along with asymptotic standard errors. These estimates would cause one to be cautious about the proposed model– standard errors are quite large, and 3 eigenvalues of the estimated ρ^Q matrix are outside the unit circle. We found small-sample standard errors much more difficult to calculate for this example, in part because the value of ρ^Q associated with a given $\hat{\pi}^{(j)}$ can have anywhere from zero to four complex eigenvalues, with eigenvalues of the $\rho^Q_{\ell\ell}$ submatrix sometimes greater than 2 in modulus. Our interpretation is that further restrictions on the interaction between the macro and latent factors could be helpful for this class of models.

$c^{\mathcal{Q}}$	0.0306	−0.0458	0	0	0
	(0.5291)	(1.1382)			
c	−0.1028	0.2414	−0.9632	−1.5301	2.4063
	(0.4951)	(0.4672)	(7.2480)	(1.4128)	(4.4009)
$\rho^{\mathcal{Q}}$	0.7725	0.2933	0.0436	−0.2138	−0.3565
	(0.2895)	(0.2801)	(1.0688)	(0.1332)	(0.3900)
	−0.3933	1.2411	0.2376	−0.0197	−0.0574
	(0.3857)	(0.3706)	(0.2437)	(0.1470)	(0.5579)
	0.2036	−0.2046	0.8579	0	0
	(0.3691)	(0.3852)	(0.1435)		
	−0.1035	0.1035	−0.0054	0.8826	−0.1926
	(0.2083)	(0.2373)	(0.5723)	(0.0672)	(0.1464)
	0.1001	−0.1415	0.0223	0.0303	0.8826
	(0.6387)	(0.6661)	(0.1215)	(0.0810)	(0.0672)
ρ	0.9461	0.2203	−0.0428	−0.0210	0.0639
	(0.0325)	(0.0508)	(0.2005)	(0.0456)	(0.1531)
	0.0002	0.8735	−0.0435	−0.0233	−0.0517
	(0.0310)	(0.0487)	(0.1618)	(0.0538)	(0.1555)
	0.0932	0.1683	0.8203	−0.0844	0.1378
	(0.3903)	(0.1686)	(0.6723)	(0.2453)	(1.0303)
	−0.0827	0.0852	−0.1110	0.8715	0.0978
	(0.1190)	(0.1295)	(0.3430)	(0.1127)	(0.2066)
	0.1220	0.0449	0.0756	0.0555	0.4728
	(0.2649)	(0.5693)	(1.0167)	(0.1468)	(0.7418)
δ_0	−0.0082				
	(0.0062)				
δ_1	6.86E-4	1.02E-3	2.03E-3	1.92E-4	7.67E-4
	(2.88E-4)	(3.03E-4)	(2.35E-3)	(1.33E-3)	(6.31E-3)
Σ_e	2.02E-4	1.87E-4	1.09E-4		
	(1.29E-5)	(1.19E-5)	(6.97E-6)		
Σ_{mm}	0.6996	0			
	(0.0448)				
	0.1174	0.6617			
	(0.0604)	(0.0424)			

Table 6: FIML estimates and asymptotic standard errors for the MF1 model.

4.4 Example 3. Macro finance model with 12 lags.

Here our data set follows Ang and Piazzesi (2003) as closely as possible, using zero-coupon bond yields with maturities of 1, 3, 12, 36 and 60 months from CRSP monthly treasury file, each divided by 1200 to quote as monthly fractional rates. We obtained two groups of monthly US macroeconomic key indicators, seasonally adjusted if applicable, from Datastream. The first group consists of various inflation measures which are based on the CPI, the PPI of finished goods, and the CRB Spot Index for commodity prices. The second group contains variables that capture real activity: the Index of Help Wanted Advertising, Unemployment Rates, the growth rate of Total Civilian Employment and the growth rate of Industrial Production. All growth rates and inflation rates are measured as the difference in logs of the monthly index value between dates t and $t-12$. We first normalized each series separately to have zero mean and unit variance, then extracted the first principal component of each group, designated the "inflation" and "real activity" indices, respectively, with each index having zero mean and unit variance by construction. The sample period for yields is from December 1952 to December 2000, and that for the macro indices is from January 1952 to December 2000. We assume that 1-, 12- and 60-month yields are priced exactly, and 3- and 36-month yields are priced with error ($N_e = 2$). We use the Ang and Piazzesi (2003) Macro Model with their additional proposed zero restrictions to illustrate minimum-chi-square estimation for an overidentified model.

The reduced-form equations (49)-(51) form 3 independent blocks. If we interpret $Y_t^m =$

f_t^m, we can write the structure of block i for $i = 1, 2, m$ as

$$\underset{(q_i \times 1)}{Y_t^i} = \underset{(q_i \times k_i)}{\Pi_i'} \underset{(k_i \times 1)}{x_{it}} + \underset{(q_i \times 1)}{u_{it}^*}$$

$$u_{it}^* \sim N(0, \Omega_i^*).$$

The information matrix for the full system of reduced-form parameters is

$$\hat{R} = \begin{bmatrix} \hat{R}_m & 0 & 0 \\ 0 & \hat{R}_1 & 0 \\ 0 & 0 & \hat{R}_2 \end{bmatrix}$$

where as in Magnus and Neudecker (1988, p. 321)

$$\hat{R}_i = \begin{bmatrix} \hat{\Omega}_i^{*-1} \otimes T^{-1} \sum_{t=1}^{T} x_{it} x_{it}^{'} & 0 \\ 0 & (1/2) D_{q_i}^{'} \left(\hat{\Omega}_i^{*-1} \otimes \hat{\Omega}_i^{*-1} \right) D_{q_i} \end{bmatrix}$$

for D_N the $N^2 \times N(N+1)/2$ duplication matrix satisfying $D_N \text{vech}(\Omega) = \text{vec}(\Omega)$.

The structural parameters Σ_e appear only in the last half of the third block, no other parameters appear in this block, and these 2 structural parameters are just-identified by the 2 diagonal elements of Ω_2^*. Thus the minimum-chi-square estimates of Σ_e are obtained immediately from the square roots of diagonal elements of $\hat{\Omega}_2^*$. The structural parameters $\rho_1, ..., \rho_{12}$ appear directly in the first block and, through ρ^Q, in the second and third blocks as well, so FIML or minimum-chi-square estimation would exploit this. However, to reduce dimensionality, we follow Ang and Piazzesi in replacing $\rho_2, ..., \rho_{12}$ where they appear in ρ^Q with

the OLS estimates $\hat{\rho}_2, ..., \hat{\rho}_{12}$. In order to try to replicate their setting as closely as possible, we also follow their procedure of imposing $\hat{\delta}_{1m}$ on the basis of OLS estimation of (23). Hence the minimum-chi-square analog to their problem is to minimize an expression of the form of (52) with

$$\hat{\pi} = \left(\left[\text{vec} \left(\hat{\Pi}_1 \right) \right]', \left[\text{vech} \left(\hat{\Omega}_1^* \right) \right]', \left[\text{vec} \left(\hat{\Pi}_2 \right) \right]' \right)' \tag{55}$$

$$\hat{R} = \begin{bmatrix} \hat{\Omega}_1^{*-1} \otimes T^{-1} \sum_{t=1}^{T} x_{1t} x_{1t}' & 0 & 0 \\ 0 & (1/2) D_3' \left(\hat{\Omega}_1^{*-1} \otimes \hat{\Omega}_1^{*-1} \right) D_3 & 0 \\ 0 & 0 & \hat{\Omega}_2^{*-1} \otimes T^{-1} \sum_{t=1}^{T} x_{2t} x_{2t}' \end{bmatrix}$$

$$x_{1t} = (1, F_{t-1}^{m\prime}, Y_{t-1}^{1\prime}, f_t^{m\prime})'$$

$$x_{2t} = (1, F_t^{m\prime}, Y_t^{1\prime})'$$

$$\hat{\Pi}_i' = \left(\sum_{t=1}^{T} Y_t^i x_{it}' \right) \left(\sum_{t=1}^{T} x_{it} x_{it}' \right)^{-1} \quad \text{for } i = 1, 2$$

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

$$\hat{\Omega}_2^* = T^{-1} \sum_{t=1}^{T} \begin{bmatrix} [\hat{u}_{2t}(1)]^2 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & [\hat{u}_{2t}(N_e)]^2 \end{bmatrix}$$

with $\hat{u}_{2t}(j)$ the jth element of $Y_t^2 - \hat{\Pi}_2' x_{2t}$.

Ang and Piazzesi also imposed a further set of restrictions on parameters, setting parameters with large standard errors as estimated in their first stage to zero. Their understanding was that the purpose of these restrictions was to improve efficiency, though we saw in Section

3.3 that some of these restrictions are in fact necessary in order to achieve identification. Our purpose here is to illustrate the minimum-chi-squared method on an overidentified structure, and we therefore attempt to estimate their final proposed structure using our method. The additional parameters that Ang and Piazzesi fixed at zero include the (2,1) and (3,1) elements of $\rho_{\ell\ell}$ (which recall was already lower triangular), the (1,2), (2,2), (3,2) and (1,3) elements of $\Lambda_{\ell\ell}$, both elements in λ_m, and the 2^{nd} and 3^{rd} elements of λ_ℓ. Our goal is then to minimize (52) with respect to the 17 remaining unknown parameters, 1 in λ_ℓ, 4 in Λ_{mm}, 5 in $\Lambda_{\ell\ell}$, 4 in $\rho_{\ell\ell}$, and 3 in $\delta_{1\ell}$.[17]

The results of this estimation for 100 different starting values are reported in Table 7. Our procedure uncovered three local minima to the objective function. The parameters we report as Local1 correspond to the values reported in Table 6 of Ang and Piazzesi. The small differences between our estimates and theirs are due to some slight differences between the data sets and the fact that, in an overidentified structure, the minimum-chi-square and maximum-likelihood estimates are not numerically identical. Our procedure establishes that the estimates reported by Ang and Piazzesi in fact represent only a local maximum of the likelihood– both the estimates we report as Local2 and Global achieve substantially higher values for the log likelihood function relative to Local1. Moreover, the differences between

[17] We made one other slight change in parameterization that may be helpful. Since $\Lambda_{\ell\ell}$ always enters either the minimum-chi-squared calculations or the original maximum likelihood estimation in the form of high powers of the matrix $\rho^Q_{\ell\ell} = \rho_{\ell\ell} - \Lambda_{\ell\ell}$, the algorithms will be better behaved numerically if the unknown elements of $\rho^Q_{\ell\ell}$ rather than those of $\Lambda_{\ell\ell}$ are taken to be the object of interest. Specifically, for this example we implemented this subject to the proposed restrictions by parameterizing

$$\rho_{\ell\ell} = \begin{bmatrix} \theta_1 & 0 & 0 \\ 0 & \theta_2 & 0 \\ 0 & \theta_3 & \theta_4 \end{bmatrix} \quad \rho^Q_{\ell\ell} = \begin{bmatrix} \theta_5 & 0 & 0 \\ \theta_6 & \theta_2 & \theta_7 \\ \theta_8 & \theta_3 & \theta_9 \end{bmatrix},$$

and then translated back in terms of the implied values for $\Lambda_{\ell\ell}$ for purposes of reporting values in Table 7.

estimates in terms of the pricing of risk are substantial. In the original reported Ang and Piazzesi estimates, an increase in inflation lowers the price of inflation risk and raises the price of output risk, whereas the values implied by Global reverse these signs. This is consistent with their finding that the prices of observable macro risk behave very differently between their Macro Model and Macro Lag Model specifications– we find they also differ substantially across alternative local maxima of the log likelihood function even within their single Macro Model specification. Note that the large prices of risk for these higher local maxima can make them easy to miss with conventional estimation and conventional starting values of zero price of risk.

	Global			Local1			Local2		
$\rho_{\ell\ell}$	0.9921	0	0	0.9918	0	0	0.9920	0	0
	0	0.9462	0	0	0.9412	0	0	0.9437	0
	0	-0.0034	0.9021	0	-0.0095	0.7712	0	-0.0032	0.9401
$\delta_{1\ell}$	1.11E-04	4.27E-04	1.98E-04	1.09E-04	4.30E-04	1.92E-04	1.22E-04	4.26E-04	1.92E-04
λ_ℓ	-0.0409	0	0	-0.0441	0	0	-0.0388	0	0
Λ_{mm}	2.8783	0.4303		-0.3430	0.1474		1.5633	0.1341	
	-6.1474	-0.8744		1.7675	-0.0607		16.0624	7.4290	
$\Lambda_{\ell\ell}$	-0.0048	0	0	-0.0045	0	0	-0.0056	0	0
	-0.0445	0	0.2910	-0.0474	0	0.2881	-0.0423	0	0.3000
	-0.0322	0	0.3687	-0.0331	0	0.2110	-0.0299	0	0.4120
χ^2	462.15			530.69			503.10		
LLF	20703			20668			20679		
Frequency	14			84			2		

Table 7: Three local minima for the chi-square objective function for the restricted MF12 specification.

Another benefit of the minimum-chi-square estimation is that the value for the objective function itself gives us an immediate test of the various overidentifying restrictions. There are 152 parameters in the reduced form vector π in (55). The 17 estimated elements of θ then leave 135 degrees of freedom. The 1% critical value for a $\chi^2(135)$ variable is 176. Thus the

observed minimum value for our objective function (462.15) provides overwhelming evidence that the restrictions imposed by the model are inconsistent with the observed data.

5 Conclusion.

We have characterized affine term structure models entirely in terms of their implications for the parameters of a vector autoregression in observed variables. In addition to helping to understand exactly what features of the data led to the particular asset-pricing parameter estimates, this perspective allows us to evaluate the identifiability of a proposed structure. We demonstrated that a failure of local or global identification can complicate traditional estimation efforts, leading researchers to impose somewhat arbitrary restrictions in order to obtain estimates and in other cases miss the true global maximum of the likelihood function. We proposed an alternative estimation philosophy based on minimum-chi-square estimation, which lets OLS find an unrestricted maximum of the likelihood surface and backs out from the OLS estimates the optimal implied values for the structural parameters of interest. Among other benefits, this approach makes it feasible in some cases to calculate small-sample standard errors, instantly know whether estimates represent a global or only a local optimum, and recognize whether a given structure is unintentionally restricting the class of possible affine term structure models.

By showing how to recognize an unidentified structure, greatly reducing the computational burden of estimation, and providing an immediate specification test of any proposed restrictions, we hope that our methods will help to make these models a more effective tool for

research in macroeconomics and finance.

References

Aït-Sahalia, Yacine and Robert L. Kimmel, "Estimating Affine Multifactor Term Structure Models Using Closed-Form Likelihood Expansions," *Journal of Financial Economics*, 2010, *98.*

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.

_ , _ , and Min Wei, "What does the yield curve tell us about GDP growth," *Journal of Econometrics*, 2006, *131*, 359–403.

_ , Sen Dong, and Monika Piazzesi, "No-arbitrage Taylor rules," 2007. National Bureau of Economic Research, Working paper no. 13448.

Bauer, Michael D., "Term premia and the news." PhD dissertation, University of California, San Diego 2009.

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.

Bekaert, Geert, Seonghoon Cho, and Antonio Moreno, "New Keynesian macroeconomics and the term structure," *Journal of Money, Credit, and Banking*, 2010, *42*, 33–62.

Chamberlain, Gary, "Multivariate Models for Panel Data," *Journal of Econometrics*, 1982, *18*, 5–46.

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*, forthcoming.

_ , Jose A. Lopez, and Glenn D. Rudebusch, "Do central bank liquidity facilities affect interbank lending rates?," 2009. Federal Reserve Bank of San Francisco, Working paper 2009-13.

_ , _ , and _ , "Inflation expectations and risk premiums in an arbitrage-free model of nominal and real bond yields," 2010. Federal Reserve Bank of San Francisco, Working paper 2008-34.

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

Collin-Dufresne, Pierre, Robert S. Goldstein, and Christopher S. Jones, "Identification of Maximal Affine Term Structure Models," *Journal of Finance*, 2008, *63* (2), 743–795.

Dai, Qiang and Kenneth J. Singleton, "Specification analysis of affine term structure models," *The Journal of Finance*, 2000, *55*, 1943–1978.

_ **and** _ , "Expectation puzzles, time-varying risk premia, and affine models of the term structure," *Journal of Financial Economics*, 2002, *63*, 415–441.

Duarte, Jefferson, "Evaluating an alternative risk preference in affine term structure models," *Review of Financial Studies*, 2004, *17*, 379–404.

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," 2009. Working Paper, Johns Hopkins University.

Duffie, Darrell and Rui Kan, "A yield-factor model of interest rates," *Mathematical Finance*, 1996, *6*, 379–406.

Fisher, Franklin M., *The Identification Problem in Econometrics*, New York: McGraw-Hill, 1966.

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

_ **and Jing Wu**, "The Effectiveness of Alternative Monetary Policy Tools in a Zero Lower Bound Environment," 2010. Working paper, University of California, San Diego.

_ **and** _ , "Testable Implications of Affine-Term-Structure Models," 2010. Working paper, University of California, San Diego.

Hordahl, Peter, Oreste Tristani, and David Vestin, "A joint econometric model of macroeconomic and term structure dynamics," *Journal of Econometrics*, 2006, *131*, 405–44.

Joslin, Scott, Kenneth J. Singleton, and Haoxiang Zhu, "A new perspective on Gaussian dynamic term structure models," *Review of Financial Studies*, forthcoming.

Kim, Don H., "Challenges in macro-finance modeling," 2008. BIS Working Paper No. 240, FEDS Working Paper No. 2008-06.

_ **and Athanasios Orphanides**, "Term structure estimation with survey data on interest rate forecasts," 2005. Federal Reserve Board, Finance and Economics Discussion Series 2005-48.

_ **and Jonathan H. Wright**, "An arbitrage-free three-factor term structure model and the recent behavior of long-term yields and distant-horizon forward rates," 2005. Federal Reserve Board, Finance and Economics Discussion Series 2005-33.

Magnus, Jan R. and Heinz Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, Ltd., 1988.

Newey, Whitney K., "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Journal of Econometrics*, 1987, *36* (3), 231–250.

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., "Identification in parametric models," *Econometrica*, 1971, *39*, 577–591.

_ , *Efficient Estimation with A Priori Information*, Yale University Press, 1973.

Rudebusch, Glenn D. 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.

Singleton, Kenneth J., *Empirical Dynamic Asset Pricing*, Princeton University Press, 2006.

Smith, Josephine M., "The term structure of money market spreads during the financial crisis." PhD dissertation, Stanford University 2010.

Vasicek, Oldrich, "An Equilibrium Characterization of the Term Structure," *Journal of Financial Economics*, 1977, *5*, 177–188.

Appendix A. Log likelihood function for the MF1 specification.

The coefficients relating Y_t^1 and Y_t^2 to macro and latent factors can be partitioned as

$$\begin{bmatrix} B_{1m} & B_{1\ell} \\ {\scriptstyle (3\times 2)} & {\scriptstyle (3\times 3)} \\ B_{2m} & B_{2\ell} \\ {\scriptstyle (3\times 2)} & {\scriptstyle (3\times 3)} \end{bmatrix} = \begin{bmatrix} b_4' \\ b_{20}' \\ b_{40}' \\ b_8' \\ b_{12}' \\ b_{28}' \end{bmatrix}$$

for b_n given by (15). The conditional density for the tth observation is then

$$f(f_t^m, Y_t | f_{t-1}^m, Y_{t-1}) = \frac{1}{|\det(J)|} f(f_t^m, f_t^\ell, u_t^e | f_{t-1}^m, f_{t-1}^\ell, u_{t-1}^e)$$

where

$$f(f_t^m, f_t^\ell, u_t^e | f_{t-1}^m, f_{t-1}^\ell, u_{t-1}^e) = f(f_t^m | f_{t-1}^\ell, f_{t-1}^m) f(f_t^\ell, | f_{t-1}^\ell, f_{t-1}^m) f(u_t^e)$$

$$f(f_t^m | f_{t-1}^\ell, f_{t-1}^m) = \phi(f_t^m; c_m + \rho_{mm} f_{t-1}^m + \rho_{m\ell} f_{t-1}^\ell, \Sigma_{mm}\Sigma_{mm}')$$

$$f(f_t^\ell | f_{t-1}^\ell, f_{t-1}^m) = \phi(f_t^\ell; c_\ell + \rho_{\ell m} f_{t-1}^m + \rho_{\ell\ell} f_{t-1}^\ell, I_{N_\ell})$$

$$f(u_t^e) = \phi(u_t^e; 0, I_{N_e})$$

$$f_t^\ell = B_{1\ell}^{-1}(Y_t^1 - A_1 - B_{1m} f_t^m)$$

$$u_t^e = \Sigma_e^{-1}(Y_t^2 - A_2 - B_{2m} f_t^m - B_{2\ell} f_t^\ell)$$

$$J = \begin{bmatrix} B_{1\ell} & 0 \\ B_{2\ell} & \Sigma_e \end{bmatrix}.$$

For the Q representation and our $N_\ell = 3$, $N_m = 2$, $N_e = 3$ example, there are 25 unknown elements in ρ, 25 in ρ^Q, 5 in c, 2 in c^Q, 5 in δ_1, 1 in δ_0, 3 in Σ_{mm}, and 3 in Σ_e. The traditional approach is to arrive at estimates of these 69 parameters by numerical maximization of

$$\mathcal{L}(\theta; Y) = \sum_{t=1}^{T} \log f(f_t^m, Y_t | f_{t-1}^m, Y_{t-1})$$

as calculated using the above formulas.

Appendix B. Log likelihood for the MF12 specification.

The P dynamics can again be represented as a special case of (1) by using the companion form $F_t = (F_t^{m\prime}, f_t^{\ell\prime})'$, $F_t^m = (f_t^{m\prime}, ..., f_{t-11}^{m\prime})'$, $c = \left(0'_{24\times 1}, c'_\ell\right)'$, and

$$\rho = \begin{bmatrix} \underset{(2\times 2)}{\rho_1} & \rho_2 & \rho_3 & \cdots & \rho_{11} & \rho_{12} & 0 \\ I_2 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & I_2 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & I_2 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \underset{(3\times 3)}{\rho_{\ell\ell}} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \underset{(2\times 2)}{\Sigma_{mm}} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \underset{(3\times 3)}{\Sigma_{\ell\ell}} \end{bmatrix}.$$

Ang and Piazzesi assumed that the risk associated with lagged macro factors is not priced and imposed the restriction in a λ representation that the values in (9) are characterized by $\lambda = \left(\lambda'_m, 0'_{22\times 1}, \lambda'_\ell\right)'$ and

$$\underset{(27\times 27)}{\Lambda} = \begin{bmatrix} \underset{(2\times 2)}{\Lambda_{mm}} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \underset{(3\times 3)}{\Lambda_{\ell\ell}} \end{bmatrix}.$$

>From (10) and (11) it follows that the parameters in (12) are given by $c^Q = \left(c_m^{Q\prime}, 0'_{22\times 1}, c_\ell^{Q\prime}\right)'$ and

$$\rho^Q = \begin{bmatrix} \underset{(2\times 2)}{\rho_1^Q} & \rho_2 & \rho_3 & \cdots & \rho_{11} & \rho_{12} & 0 \\ I_2 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & I_2 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & I_2 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \underset{(3\times 3)}{\rho_{\ell\ell}^Q} \end{bmatrix}.$$

Ang and Piazzesi used $N_\ell = 3$ and $N_e = 2$, assuming that the 1-, 12-, and 60-month yields were priced without error, while the 3- and 36-month yields were priced with error, so that the the B matrices can be written in partitioned form as

$$\begin{bmatrix} B_{1m}^{(0)} & B_{1m}^{(1)} & B_{1\ell} \\ (3\times 2) & (3\times 22) & (3\times 3) \\ B_{2m}^{(0)} & B_{2m}^{(1)} & B_{2\ell} \\ (2\times 2) & (2\times 22) & (2\times 3) \end{bmatrix} = \begin{bmatrix} b_1' \\ b_{12}' \\ b_{60}' \\ b_3' \\ b_{36}' \end{bmatrix}$$

where for example $B_{1m}^{(1)}$ are the coefficients relating the observed yields to 11 lags of the 2 macro factors.

The conditional density for this case is then

$$f(f_t^m, Y_t | F_{t-1}^m, Y_{t-1}) = \frac{1}{|\det(J)|} f(f_t^m, f_t^\ell, u_t^e | F_{t-1}^m, f_{t-1}^\ell, u_{t-1}^e)$$

$$f(f_t^m, f_t^\ell, u_t^e | F_{t-1}^m, f_{t-1}^\ell, u_{t-1}^e) = f(f_t^m | F_{t-1}^m) f(f_t^\ell | f_{t-1}^\ell) f(u_t^e)$$

$$f(f_t^m | F_{t-1}^m) = \phi(f_t^m; \rho_1 f_{t-1}^m + \rho_2 f_{t-2}^m + \cdots + \rho_{12} f_{t-12}^m, \Sigma_{mm}\Sigma_{mm}')$$

$$f(f_t^\ell | f_{t-1}^\ell) = \phi(f_t^\ell; \rho_{\ell\ell} f_{t-1}^\ell, I_{N_\ell})$$

$$f(u_t^e) = \phi(u_t^e; 0, I_{N_e})$$

$$f_t^\ell = B_{1\ell}^{-1}\left(Y_t^1 - A_1 - \begin{bmatrix} B_{1m}^{(0)} & B_{1m}^{(1)} \end{bmatrix} F_t^m\right)$$

$$u_t^e = \Sigma_e^{-1}\left(Y_t^2 - A_2 - \begin{bmatrix} B_{2m}^{(0)} & B_{2m}^{(1)} \end{bmatrix} F_t^m - B_{2\ell} f_t^\ell\right)$$

$$J = \begin{bmatrix} B_{1\ell} & 0 \\ B_{2\ell} & \Sigma_e \end{bmatrix}.$$

Appendix C. Proof of Proposition 1.

Write

$$H = \begin{bmatrix} u & x \\ v & y \end{bmatrix}.$$

Since columns of H have unit length, without loss of generality we can write $(u, v) = (\cos\theta, \sin\theta)$ for some $\theta \in [-\pi, \pi]$. The second column of H is also a point on the unit circle, for which orthogonality with the first column also requires it to be located on the line $ux + vy = 0$, with the two solutions $x = -v$, $y = u$ and $x = v$, $y = -u$. Thus the set of orthogonal (2×2) matrices can be represented as either rotations

$$H_1(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{C.1}$$

or reflections

$$H_2(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}. \tag{C.2}$$

The condition that the $(1,2)$ element of $H_1(\theta)\rho H_1(\theta)'$ be zero requires

$$(\rho_{11}^Q - \rho_{22}^Q)\sin\theta\cos\theta - \rho_{21}^Q \sin^2\theta = 0.$$

One way this could happen is if $\sin\theta = 0$. But this would imply either $H_1(-\pi/2) = -I_2$, violating the sign requirement $H\delta_1 \geq 0$, or else the identity transformation$H_1(\pi/2) = I_2$. Hence the condition of interest is

$$(\rho_{11}^Q - \rho_{22}^Q)\cos\theta - \rho_{21}^Q \sin\theta = 0. \tag{C.3}$$

If θ_1 satisfies condition (C.3), then one can show

$$H_1(\theta_1)\rho^Q H_1(\theta_1)' = \begin{bmatrix} \rho_{22}^Q & 0 \\ \rho_{21}^Q & \rho_{11}^Q \end{bmatrix}.$$

Alternatively for $H_2(\theta)$ we have the requirement

$$(\rho_{11}^Q - \rho_{22}^Q)\sin\theta\cos\theta + \rho_{21}^Q \sin^2\theta = 0$$

for which the solution $\sin\theta = 0$ would violate $H_2(\theta)\delta_1 \geq 0$, leaving the sole condition

$$(\rho_{11}^Q - \rho_{22}^Q)\cos\theta + \rho_{21}^Q \sin\theta = 0. \tag{C.4}$$

For any θ_2 satisfying (C.4),

$$H_2(\theta_2)\rho^Q H_2(\theta_2)' = \begin{bmatrix} \rho_{22}^Q & 0 \\ -\rho_{21}^Q & \rho_{11}^Q \end{bmatrix}.$$

Now consider the nonnegativity condition. Since $\cot\theta$ is monotonic on $(0,\pi)$ and repeats the pattern on $(-\pi,0)$, there are two values $\theta \in [-\pi,\pi]$ satisfying (C.3). We denote the first by $\theta_1 \in [0,\pi]$, in which case the second is given by $\theta_1 - \pi$. The two solutions to (C.4) can then be written as $-\theta_1$ and $-\theta_1 + \pi$. We are then looking at 4 possible transformations:

$$H_1(\theta_1)\delta_1 = \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{bmatrix}\begin{bmatrix} \delta_{11} \\ \delta_{12} \end{bmatrix} = \begin{bmatrix} \delta_{11}\cos\theta_1 - \delta_{12}\sin\theta_1 \\ \delta_{11}\sin\theta_1 + \delta_{12}\cos\theta_1 \end{bmatrix} \equiv \begin{bmatrix} \delta_{11}^* \\ \delta_{12}^* \end{bmatrix}$$

$$H_1(\theta_1 - \pi)\delta_1 = \begin{bmatrix} -\cos\theta_1 & \sin\theta_1 \\ -\sin\theta_1 & -\cos\theta_1 \end{bmatrix}\begin{bmatrix} \delta_{11} \\ \delta_{12} \end{bmatrix} = \begin{bmatrix} -\delta_{11}^* \\ -\delta_{12}^* \end{bmatrix}$$

$$H_2(-\theta_1)\delta_1 = \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 \\ -\sin\theta_1 & -\cos\theta_1 \end{bmatrix}\begin{bmatrix} \delta_{11} \\ \delta_{12} \end{bmatrix} = \begin{bmatrix} \delta_{11}^* \\ -\delta_{12}^* \end{bmatrix}$$

$$H_2(-\theta_1 + \pi) = \begin{bmatrix} -\cos\theta_1 & \sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{bmatrix}\begin{bmatrix} \delta_{11} \\ \delta_{12} \end{bmatrix} = \begin{bmatrix} -\delta_{11}^* \\ \delta_{12}^* \end{bmatrix}.$$

Apart from the knife-edge condition $\delta_{11}^* = 0$ or $\delta_{12}^* = 0$ (which would require a particular relation between the elements of the original ρ^Q and δ_1), one and only one of the above four vectors would have both elements positive, and this matrix produces $H\rho^Q H'$ of one of the two specified forms.

For $N_\ell > 2$, one can construct a family of such orthogonal matrices, for example using a matrix like

$$H(\theta) = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}$$

for θ satisfying $\rho_{31}^Q \sin\theta = (\rho_{11}^Q - \rho_{33}^Q)\cos\theta$, which swaps the (1,1) and (3,3) elements of ρ^Q. Exactly one of the 4 possible matrices performing this swap will preserve positive $H\delta_1$. There are N_ℓ choices for the value one can put into the (1,1) element as a result of such swaps, $N_\ell - 1$ remaining choices for ρ_{22}^Q, or a total of $N_\ell!$ permutations.

Appendix D. Proof of Proposition 2.

Consider first rotations $H_1(\theta)$ as specified in (C.1). The (1,1) element of $\Upsilon = H_1(\theta)\rho^Q[H_1(\theta)]'$ is seen to be

$$h_1(\theta) = \rho_{11}^Q \cos^2\theta - (\rho_{21}^Q + \rho_{12}^Q)\cos\theta\sin\theta + \rho_{22}^Q \sin^2\theta. \tag{D.1}$$

We claim first that there exists a $\theta \in [0, \pi/2]$ such that $h_1(\theta)$ equals $(\rho_{11}^Q + \rho_{22}^Q)/2$. To see this, note that at $\theta = 0$, the value of $h_1(\theta)$ is ρ_{11}^Q, whereas at $\theta = \pi/2$, it is instead equal to ρ_{22}^Q. Since $h_1(\theta)$ is continuous in θ, there exists a value θ_1 such that $h_1(\theta_1)$ is exactly halfway between ρ_{11}^Q and ρ_{22}^Q.

Notice next that the eigenvalues of $\Upsilon = H\rho^Q H'$ are identical to those of ρ^Q, and hence the trace of Υ (which is the sum of the eigenvalues) is the same as the trace of ρ^Q:

$$\Upsilon_{11} + \Upsilon_{22} = \rho_{11}^Q + \rho_{22}^Q.$$

Thus since $\Upsilon_{11} = (\rho_{11}^Q + \rho_{22}^Q)/2$, then also $\Upsilon_{22} = (\rho_{11}^Q + \rho_{22}^Q)/2$. Hence $H_1(\theta_1)\rho^Q[H_1(\theta_1)]'$ is of the desired form with elements along the principal diagonal equal to each other. As in the proof of Proposition 1, $H_1(\theta_1 - \pi)$ is the other rotation that works.

Alternatively, H could be a reflection matrix $H_2(\theta)$ as in (C.2), for which the (1,1) element of $H_2(\theta)\rho^Q[H_2(\theta)]'$ is found to be:

$$\rho_{11}^Q \cos^2\theta + (\rho_{21}^Q + \rho_{12}^Q)\cos\theta\sin\theta + \rho_{22}^Q \sin^2\theta \tag{D.2}$$

This turns out to equal $(\rho_{11}^Q + \rho_{22}^Q)/2$ at $\theta_2 = -\theta_1$ and $\theta_2 = -\theta_1 + \pi$. As in the proof of Proposition 1, in the absence of knfe-edge conditions on δ_1, exactly one of the transformations $H_1(\theta_1)$, $H_1(\theta_1 - \pi)$, $H_2(-\theta_1)$, $H_2(-\theta_1 + \pi)$ preserves positivity of $H\delta_1$, establishing existence.

For uniqueness, suppose we have found a transformation $H\rho^Q H' = \Upsilon$ of the desired form. Then any alternative transformation $H^*\rho^Q H^{*\prime}$ can equivalently be written as $\tilde{H}\Upsilon\tilde{H}'$ for $\tilde{H}H = H^*$. Hence the result will be established if we can show that the only transformations $\tilde{H}\Upsilon\tilde{H}'$ that keep the diagonal elements equal to each other and also satisfy $\tilde{H}\delta_1 \geq 0$ are the identity and transposition. Since $a = \Upsilon_{11} = \Upsilon_{22}$ and since the transformation preserves eigenvalues,

we know that if the (1,1) and (2,2) elements of $\tilde{H}\Upsilon\tilde{H}'$ are equal to each other, each must again be the value a. Thus if $\tilde{H} = H_1(\theta)$ for some θ, we require as in (D.1) that

$$a\cos^2\theta - (\Upsilon_{21} + \Upsilon_{12})\cos\theta\sin\theta + a\sin^2\theta = a$$

which can only be true if

$$(\Upsilon_{21} + \Upsilon_{12})\cos\theta\sin\theta = 0. \tag{D.3}$$

This requires either $\cos\theta = 0$, $\sin\theta = 0$, or $\Upsilon_{21} = -\Upsilon_{12}$. For $\cos\theta = 0$, $H_1(\theta)\delta_1$ would violate the nonnegativity condition, while $\sin\theta = 0$ corresponds to $H_1(\theta) = \pm I_2$. Finally, if $\Upsilon_{21} = -\Upsilon_{12}$, one can verify that $H_1(\theta)\Upsilon[H_2(\theta)]' = \Upsilon$ for all θ. Alternatively, for reflections applied to a matrix Υ for which $a = \Upsilon_{11} = \Upsilon_{22}$, we see as in (D.2) that $a\cos^2\theta + (\Upsilon_{21} + \Upsilon_{12})\cos\theta\sin\theta + a\sin^2\theta = a$, which again can only hold for θ satisfying (D.3). In this case, $\sin\theta = 0$ is ruled out by the constraint $H_2(\theta)\delta_1 \geq 0$, but for $\cos\theta = 0$ we have

$$H_2(\pi/2) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and } H_2(-\pi/2) = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}.$$

Both of these give $H\Upsilon H' = \Upsilon'$ but only $H_2(\pi/2)\delta_1 > 0$. Finally, when $\Upsilon_{21} = -\Upsilon_{12}$, then $H_2(\theta)\Upsilon[H_2(\theta)]' = \Upsilon'$ for any θ. Thus the only transformation $\tilde{H}\Upsilon\tilde{H}'$ that preserves equality of diagonal elements is transposition, as claimed.

Appendix E. Asymptotic standard errors of MLE.

Here we demonstrate that under the usual regularity conditions,

$$E\left[\left.\frac{\partial^2\mathcal{L}(\pi(\theta);Y)}{\partial\theta\partial\theta'}\right|_{\theta=\theta_0}\right] = -T\Gamma' R\Gamma$$

for

$$\Gamma = \left[\left.\frac{\partial\pi(\theta)}{\partial\theta'}\right|_{\theta=\theta_0}\right]$$

$$R = -T^{-1}E\left[\left.\frac{\partial^2\mathcal{L}(\pi;Y)}{\partial\pi\partial\pi'}\right|_{\pi=\pi_0}\right].$$

Note

$$\frac{\partial\mathcal{L}(\pi(\theta);Y)}{\partial\theta'} = \begin{bmatrix} \frac{\partial\mathcal{L}(\pi)}{\partial\pi_1} & \cdots & \frac{\partial\mathcal{L}(\pi)}{\partial\pi_q} \end{bmatrix} \begin{bmatrix} \frac{\partial\pi_1}{\partial\theta_1} & \cdots & \frac{\partial\pi_1}{\partial\theta_N} \\ \vdots & \vdots & \vdots \\ \frac{\partial\pi_q}{\partial\theta_1} & \cdots & \frac{\partial\pi_q}{\partial\theta_N} \end{bmatrix}$$

$$\frac{\partial^2 \mathcal{L}(\pi(\theta);Y)}{\partial\theta_i\partial\theta'} = \begin{bmatrix} \frac{\partial\pi_1}{\partial\theta_i} & \cdots & \frac{\partial\pi_q}{\partial\theta_i} \end{bmatrix} \begin{bmatrix} \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_1\partial\pi_1} & \cdots & \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_1\partial\pi_q} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_q\partial\pi_1} & \cdots & \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_q\partial\pi_q} \end{bmatrix} \begin{bmatrix} \frac{\partial\pi_1}{\partial\theta_1} & \cdots & \frac{\partial\pi_1}{\partial\theta_N} \\ \vdots & \vdots & \vdots \\ \frac{\partial\pi_q}{\partial\theta_1} & \cdots & \frac{\partial\pi_q}{\partial\theta_N} \end{bmatrix} \tag{E.1}$$
$$+ \begin{bmatrix} \frac{\partial\mathcal{L}(\pi)}{\partial\pi_1} & \cdots & \frac{\partial\mathcal{L}(\pi)}{\partial\pi_q} \end{bmatrix} \begin{bmatrix} \frac{\partial^2\pi_1}{\partial\theta_1\partial\theta_i} & \cdots & \frac{\partial^2\pi_1}{\partial\theta_N\partial\theta_i} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2\pi_q}{\partial\theta_1\partial\theta_i} & \cdots & \frac{\partial^2\pi_q}{\partial\theta_N\partial\theta_i} \end{bmatrix}.$$

Evaluate (E.1) at $\theta = \theta_0$, take expectations with respect to the distribution of Y, and use the fact that Γ is not a function of Y:

$$E\left[\frac{\partial^2 \mathcal{L}(\pi(\theta);Y)}{\partial\theta_i\partial\theta'}\bigg|_{\theta=\theta_0}\right] = e_i'\Gamma' E \begin{bmatrix} \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_1\partial\pi_1}\Big|_{\pi=\pi_0} & \cdots & \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_1\partial\pi_q}\Big|_{\pi=\pi_0} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_q\partial\pi_1}\Big|_{\pi=\pi_0} & \cdots & \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_q\partial\pi_q}\Big|_{\pi=\pi_0} \end{bmatrix} \Gamma \tag{E.2}$$
$$+ \left\{ E\begin{bmatrix} \frac{\partial\mathcal{L}(\pi)}{\partial\pi_1}\Big|_{\pi=\pi_0} & \cdots & \frac{\partial\mathcal{L}(\pi)}{\partial\pi_q}\Big|_{\pi=\pi_0} \end{bmatrix} \right\} \begin{bmatrix} \frac{\partial^2\pi_1}{\partial\theta_1\partial\theta_i}\Big|_{\theta=\theta_0} & \cdots & \frac{\partial^2\pi_1}{\partial\theta_N\partial\theta_i}\Big|_{\theta=\theta_0} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2\pi_q}{\partial\theta_1\partial\theta_i}\Big|_{\theta=\theta_0} & \cdots & \frac{\partial^2\pi_q}{\partial\theta_N\partial\theta_i}\Big|_{\theta=\theta_0} \end{bmatrix}.$$

But the usual regularity conditions imply $E\left\{\partial\mathcal{L}(\pi)/\partial\pi_j|_{\pi=\pi_0}\right\} = 0$, so the second term in (E.2) vanishes. Stacking the row vectors represented by the first term into a matrix produces

$$E\left[\frac{\partial^2 \mathcal{L}(\pi(\theta);Y)}{\partial\theta\partial\theta'}\bigg|_{\theta=\theta_0}\right] = \Gamma' E \begin{bmatrix} \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_1\partial\pi_1}\Big|_{\pi=\pi_0} & \cdots & \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_1\partial\pi_q}\Big|_{\pi=\pi_0} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_q\partial\pi_1}\Big|_{\pi=\pi_0} & \cdots & \frac{\partial^2\mathcal{L}(\pi)}{\partial\pi_q\partial\pi_q}\Big|_{\pi=\pi_0} \end{bmatrix} \Gamma$$

as claimed.