Real-Time Measurement of Business Conditions*

S. Borağan Aruoba† Francis X. Diebold Chiara Scotti

University of Maryland University of Pennsylvania and NBER Federal Reserve Board

April 13, 2007

Abstract

We construct a framework for measuring economic activity in real time (e.g., minute-by-minute), using a variety of stock and flow data observed at mixed frequencies. Specifically, we propose a dynamic factor model that permits exact filtering. We explore the efficacy of our methods both in simulation environments and in a sequence of progressively richer empirical examples.

Key Words: Business cycle, Expansion, Recession, State space model, Macroeconomic forecasting

JEL Codes: E32, E37, C01, C22

*For helpful discussion we thank Alexi Onatski, Jon Faust, Jonathan Wright, Martin Evans and Eric Ghysels. For research support we thank the National Science Foundation. The usual disclaimer applies.

†Corresponding author. Department of Economics, University of Maryland, College Park, MD 20742. aruoba@econ.umd.edu

1 Introduction

Aggregate business conditions are of central importance in the business, finance, and policy communities, worldwide, and huge resources are devoted to assessment of the continuously-evolving state of the real economy. Literally thousands of newspapers, newsletters, television shows, and blogs, not to mention armies of employees in manufacturing and service industries, including the financial services industries, central banks, government and non-government organizations, grapple daily with the real-time measurement and forecasting of evolving business conditions.

Against this background, we will propose and illustrate a framework for real-time business conditions assessment in a systematic, replicable, and statistically optimal manner. Our framework has four key parts.

Part 1. *We work with a dynamic factor model, treating business conditions as an unobserved variable, related to observed indicators.* The appeal of latency of business conditions comes from its close coherence with economic theory, which emphasizes that the business cycle is not about any single variable, whether GDP, industrial production, sales, employment, or anything else. Rather, the business cycle is about the dynamics and interactions ("co-movements") of many variables, as forcefully argued by Lucas (1977) among many others.

Treating business conditions as latent is also a venerable tradition in empirical business cycle analysis, ranging from the earliest work to the most recent, and from the statistically informal the statistically formal. On the informal side, latency of business conditions is central to many approaches, from the classic early work of Burns and Mitchell (1946) to the recent workings of the NBER business cycle dating committee, as described for example by Hall *et al.* (2003). On the formal side, latency of business conditions is central to the popular dynamic factor framework, whether from the "small data" perspective of Geweke (1977), Sargent and Sims (1977), Stock and Watson (1989, 1991), and Diebold and Rudebusch (1996), or the more recent "large data" perspective of Stock and Watson (2002) and Forni, Hallin, Lippi and Reichlin (2000).[1]

Part 2. *We explicitly incorporate business conditions indicators measured at different frequencies.* Important business conditions indicators do in fact arrive at a variety of frequencies, including quarterly (e.g., GDP), monthly (e.g., industrial production), weekly (employment), and continuously (e.g., asset prices), and we want to be able to incorporate all

[1] For definition and discussion of small-data vs. large-data dynamic factor modeling, see Diebold (2003).

of them, to provide continuously-updated assessments in real time.

Part 3. We explicitly incorporate a continuously-evolving indicator. Given that our goal is to track the evolution of real activity in real time, it is crucial to incorporate (or at least not exclude from the outset) the real-time information flow associated with continuously-evolving indicators, such as the yield curve. For practical purposes we equate "continuously-evolving" with "daily," but intra-day information could be used as well.

Part 4. We extract and forecast latent business conditions using linear yet statistically optimal procedures, which involve no approximations. The appeal of exact as opposed to approximate procedures is obvious, but achieving exact optimality is not trivial and has proved elusive in the literature, due to complications arising from temporal aggregation of stocks vs. flows in systems with mixed-frequency data.

Related to our concerns and framework is a small but nevertheless significant literature, including Stock and Watson (1989, 1991), Mariano and Murasawa (2003), Proietti and Moauro (2006), and Evans (2005). Our contribution, however, differs from all of them.[2]

Stock and Watson (1989, 1991) work in a dynamic factor framework with exact linear filtering, but they don't consider data at different frequencies or at high frequencies. We include data at different and high frequencies, while still achieving exact linear filtering. This turns out to be a non-trivial task, requiring an original modeling approach.

Mariano and Murasawa (2003) work in a dynamic factor framework and consider data at different frequencies, but not high frequencies, and their filtering algorithm is not exact. In particular, they invoke an approximation essentially equivalent to assuming that the log of a sum equals the sum of the logs.

Proietti and Moauro (2006) work in the Mariano-Murasawa framework and are able to avoid the Mariano-Murasawa approximation, but only at the cost of moving to an extended Kalman filter, which is more tedious and involves approximations of its own.

Evans (2005) does not use a dynamic factor framework and does not use high-frequency data. Instead, he equates business conditions with GDP growth, and he uses state space methods to estimate daily GDP growth using data on preliminary, advanced and final releases of GDP, as well as a variety of other macroeconomic variables.

We proceed as follows. In section 2 we provide a detailed statement of our methodological framework, covering the state space formulation with missing data, optimal filtering and smoothing, and estimation. In section 3 we report on two pilot exercises, one based on

[2]Other related and noteworthy contributions include Shen (1996), Abeysinghe (2000), Altissimo*et al.*(2002), Liu and Hall (2001), McGuckin, Ozyildirim and Zarnowitz (2003), and Ghysels, Santa Clara and Valkanov (2004).

simulated data and one on real data, which let us illustrate our methods and assess their efficacy. In section 4 we report the results of a four-indicator system, containing quarterly GDP, monthly employment, weekly initial claims, and the daily yield curve term premium. In section 5 we conclude and offer directions for future research.

2 Methodology

Here we propose a state space macroeconomic model with an ultra-high base observational frequency, treating specification, estimation, state extraction and state prediction. Our framework facilitates exactly optimal filtering and forecasting, which we achieve throughout.

2.1 Missing Observations and Temporal Aggregation

We assume that the state of the economy evolves at a very high frequency; without loss of generality, call it "daily."[3] Similarly, we assume that all economic and financial variables evolve daily, although many are not *observed* daily. For example, we view an end-of-year wealth variable as observed each December 31, and as "missing data" for every other day of the year.

Let y_t^i denote a daily economic or financial variable, and let $\tilde{y}_t^i$ denote the same variable observed at a lower frequency (without loss of generality, call it "tilde"). The relationship between $\tilde{y}_t^i$ and y_t^i depends on whether y_t^i is a stock or flow variable. In the case of a stock variable, which by definition is a point-in-time snapshot, we have:

$$\tilde{y}_t^i = \begin{cases} y_t^i & \text{if } y_t^i \text{ is observed} \\ NA & \text{otherwise,} \end{cases}$$

where NA denotes missing data. In the case of a flow variable, the lower-frequency observations of which are functions of current and past daily observations, we have

$$\tilde{y}_t^i = \begin{cases} f(y_t^i, y_{t-1}^i, ..., y_{t-D_i}^i) & \text{if } y_t^i \text{ is observed} \\ NA & \text{otherwise,} \end{cases}$$

where D_i denotes the relevant number of days for the temporal aggregation. For ease of exposition we assume for now that D_i is fixed, but in our subsequent implementation and

[3] In our subsequent empirical work, we will indeed use a daily base observational frequency, but much higher frequencies such as second-by-second could be used if desired.

empirical work we allow for time-varying D_i, which allows us to accommodate, for example, the fact that some months have 28 days, some have 29, some have 30, and some have 31.

Satisfactory treatment of temporal aggregation remains elusive in the literature. Most authors work in logarithms and are effectively forced into the unappealing "approximation" that the log of a sum equals the sum of the logs. Mariano and Murasawa (2003), for example, assume that quarterly GDP is the geometric average of the intra-quarter monthly GDP's. Similarly, Evans (2005) assumes that the quarterly GDP growth rate is the sum of the intra-quarter daily growth rates. Proietti and Moauro (2004) use an extended Kalman filter in conjunction with a linear-Gaussian approximating model, but that approach involves a significant approximation as well.

In contrast, our framework permits exact aggregation. We work in levels, so that flow variables aggregate linearly and exactly. Specifically, we model the levels of all observed variables as stationary deviations from polynomial trends of arbitrary order. The result is a linear state space system for which the standard Kalman filter is optimal, as we now describe in detail.

2.2 State Space Formulation

We assume that underlying business conditions x_t evolve daily with AR(p) dynamics,

$$x_t = \rho_1 x_{t-1} + ... + \rho_p x_{t-p} + v_t, \tag{1}$$

where v_t is a white noise innovation with unit variance.[4] We are interested in tracking and forecasting real activity, so we use a single-factor model; that is, x_t is a scalar, as for example in Stock and Watson (1989). Additional factors could of course be introduced to track, for example, wage/price developments.

We assume that all economic variables y_t^i evolve daily, although they are not necessarily observed daily. Except when y_t^i is observed daily, a case which we treat seperately below, we assume that y_t^i depends linearly on x_t and possibly also various exogenous variables and/or lags of y_t^i:

$$y_t^i = c_i + \beta_i x_t + \delta_{i1} w_t^1 + ... + \delta_{ik} w_t^k + \gamma_{i1} y_{t-D_i}^i + ... + \gamma_{in} y_{t-nD_i}^i + \varepsilon_t^i, \tag{2}$$

where the w are exogenous variables, we include n lags of the dependent variable, and the

[4] As is well-known, identification of factor models requires normalization either on a factor loading or on the factor variance, and we choose to normalize the factor variance to unity.

ε_t^i are contemporaneously and serially uncorrelated innovations. Notice that we introduce lags of the dependent variable y_t^i in multiples of D_i, because the persistence in y_t^i is actually linked to the lower (tilde) observational frequency of $\tilde{y}_t^i$. Persistence modeled only in the higher daily frequency would be inadequate, as it would decay too quickly. We use (2) as the measurement equation for all (non-daily) stock variables.

Temporal aggregation in our framework is very simple: flow variables observed at a tilde frequency lower than daily are the sums of the corresponding daily variables,

$$\tilde{y}_t^i = \begin{cases} \sum_{j=0}^{D_i-1} y_{t-j}^i & \text{if } y_t^i \text{ is observed} \\ NA & \text{otherwise.} \end{cases}$$

The relationship between an observed flow variable and the factor then follows from (2),

$$\tilde{y}_t^i = \begin{cases} \sum_{j=0}^{D_i-1} c_i + \beta_i \sum_{j=0}^{D_i-1} x_{t-j}^i + \delta_{i1} \sum_{j=0}^{D_i-1} w_{t-j}^1 + ... + \delta_{ik} \sum_{j=0}^{D_i-1} w_{t-j}^k \\ \quad +\gamma_{i1} \sum_{j=0}^{D_i-1} y_{t-D_i-j}^i + ... + \gamma_{in} \sum_{j=0}^{D_i-1} y_{t-nD_i-j}^i + \varepsilon_t^{*i} & \text{if } y_t^i \text{ is observed} \\ NA & \text{otherwise,} \end{cases} \tag{3}$$

where $\sum_{j=0}^{D_i-1} y_{t-D_i-j}^i$ is by definition the observed flow variable one period ago ($\tilde{y}_{t-D_i}^i$), and ε_t^{*i} is the sum of the ε_t^i over the tilde period. Note that although ε_t^{*i} follows a serially correlated moving average process of order $D_i - 1$ at the daily frequency, it nevertheless remains white noise when observed at the tilde frequency, due to the cutoff in the autocorrelation function of an $MA(D_i - 1)$ process at displacement $D_i - 1$. Hence we will appropriately treat ε_t^{*i} as white noise in what follows and we have $var\left(\varepsilon_t^{*i}\right) = Dvar\left(\varepsilon_t^i\right)$.

The exogenous variables w_t^j are the key to handling trend. In particular, in the important special case where the w_t^j are simply deterministic polynomial trend terms (t, t^2 and so on), we have that

$$\sum_{j=0}^{D_i-1} \left[c_i + \delta_{i1} (t-j) + ... + \delta_{ik} (t-j)^k \right] \equiv c_i^* + \delta_{i1}^* t + ... + \delta_{ik}^* t^k, \tag{4}$$

which yields

$$\tilde{y}_t^i = \begin{cases} c_i^* + \beta_i \sum_{j=0}^{D_i-1} x_{t-j}^i + \delta_{i1}^* t + ... + \delta_{ik}^* t^k + \gamma_{i1}\tilde{y}_{t-D_i}^i + ... + \gamma_{in}\tilde{y}_{t-nD_i}^i + \varepsilon_t^{*i} & \text{if } y_t^i \text{ is observed} \\ NA & \text{otherwise.} \end{cases} \tag{5}$$

In the appendix we derive the mapping between $(c, \delta_1, \delta_2, \delta_3)$ and $(c^*, \delta_1^*, \delta_2^*, \delta_3^*)$ for the first 3 trend polynomials. In our implementation, for numerical stability, we use $t/1000$, $(t/1000)^2$ and $(t/1000)^3$ instead of simply t, t^2 and t^3.[5] We use (5) as the measurement equation for all (non-daily) flow variables.

Finally, we treat variables observed at daily frequency differently, allowing them to depend on a distributed lag of the state. To promote parsimony, we use a polynomial distributed lag (PDL) specification. Specifically, the measurement equation for a daily variable is

$$y_t^i = c_i + \beta_i^0 x_t + \beta_i^1 x_{t-1} + ... + \beta_i^{\tilde{D}} x_{t-\tilde{D}} + \delta_{i1} w_t^1 + ... + \delta_{ik} w_t^k + \gamma_{i1} y_{t-1}^i + ... + \gamma_{in} y_{t-n}^i + \varepsilon_t^i, \tag{6}$$

where the elements of $\{\beta_i^j\}_{j=0}^{\tilde{D}}$ follow a low-ordered polynomial.[6] In our subsequent empirical implementation we use a third-order polynomial.

This completes the specification of our model, which has a natural state space form, to which we now turn.

2.3 Initialization, Filtering and Smoothing

Assembling the discussion thus far, the state space representation of our model is

$$\begin{aligned} y_t &= Z_t \alpha_t + \Gamma w_t + \varepsilon_t \\ \alpha_{t+1} &= T\alpha_t + R\eta_t \\ \varepsilon_t \sim (0, H)\,, \eta_t &\sim (0, Q)\,, \end{aligned} \tag{7}$$

where y_t is an $N \times 1$ vector of observed variables, α_t is an $m \times 1$ vector of state variables, w_t is a $e \times 1$ vector of exogenous variables, and ε_t and η_t are vectors of measurement and

[5] This is simply a normalization and does not affect the other parameters of interest or the log-likelihood. We impose it because in our subsequent empirical work we have over 16,000 daily observations, in which case t^3 can be very large, which might create numerical problems.

[6] Because we assume that daily frequency is the highest available, we can treat flow and stock variables identically when they are observed daily.

transition shocks. The vector w_t includes an entry of unity for the constant, k trend terms and $N \times n$ lagged dependent variables, n for each of the n elements of the y_t vector. The exact structure of these vectors will vary across the different setups we consider below. The observed data vector y_t will have many missing values, reflecting those variables observed at a frequency lower than daily, as well as missing daily data due to holidays. At a minimum, the state vector α_t will include p lags of x_t, as implied by (1). Moreover, because the presence of flow variables requires a state vector containing all lags of x_t inside the aggregation period, in practice the dimension of α_t will be much greater than p. The system parameter matrices T, R and Q, are constant, while Z, Γ and H are not, because of the variation in the number of days in a quarter or month (D_i for each i). Time-varying matrices pose no problem for the Kalman filter.

Once the model is cast in state space form, we can apply the standard Kalman filter and smoother, because our setup is a stationary one. Although these are standard and do not require explanation, we state the algorithm here for contrast with the modified algorithm that accounts for missing values, which we present subsequently. For now, assume that there are no missing observations in Y_t for all t. Denote $\{Y_1, ..., Y_t\}$ by $\mathcal{Y}_t$ for $t = 1, ..., \mathcal{T}$, where $\mathcal{T}$ denotes the last time-series observation. For given parameters, we initialize the Kalman filter using $\alpha_1 \sim N(a_1, P_1)$ where $a_1 = 0_{m\times 1}$ and P_1 solves

$$(I - T \otimes T)\, vec(P_1) = vec(RQR'). \tag{8}$$

Given a_1 and P_1, for $t = 1, ..., \mathcal{T}$, we use the contemporaneous Kalman filtering equations, which incorporate the computation of the state vector estimate and its associated covariance matrix, denoted by $a_{t|t}$ and $P_{t|t}$. [7] Given $a_t \equiv E(\alpha_t|\mathcal{Y}_{t-1})$ and $P_t = var(\alpha_t|\mathcal{Y}_{t-1})$ the prediction equations that produce a_{t+1} and P_{t+1} are

$$\begin{aligned}
v_t &= Y_t - Z_t a_t - \Lambda X_t && (9)\\
F_t &= Z_t P_t Z_t' + H && (10)\\
a_{t|t} &= a_t + P_t Z_t' F_t^{-1} v_t && (11)\\
P_{t|t} &= P_t - P_t Z_t' F_t^{-1} Z_t P_t' && (12)\\
a_{t+1} &= T a_{t|t} && (13)\\
P_{t+1} &= T P_{t|t} T' + RQR'. && (14)
\end{aligned}$$

[7] We find that using this version of the filter improves the efficiency of the algorithm. See Durbin and Koopman (2001) for details.

Given a set of parameters and the estimates $\{a_t, P_t\}_{t=1}^{\mathcal{T}}$ given those parameters, we compute the log likelihood using the prediction error decomposition,

$$\log L = -\frac{1}{2}\sum_{t=1}^{\mathcal{T}}\left[p\log 2\pi + \left(\log|F_t| + v_t' F_t^{-1} v_t\right)\right]. \tag{15}$$

The Kalman smoother computes the conditional expectation of the state vector and its covariance matrix using all the information in the data set, which we denote by $\hat{\alpha}_t \equiv E(\alpha_t|\mathcal{Y}_{\mathcal{T}})$ and $V_t \equiv var(\alpha_t|\mathcal{Y}_{\mathcal{T}})$ for $t = 1, ..., \mathcal{T}$. The Kalman smoother recursions start from $t = \mathcal{T}$ and work backward. The vector r_t is a weighed average of the innovations v_t that happen after period t with the variance matrix N_t. We initialize the smoother with $r_{\mathcal{T}} = 0_{m\times 1}$ and $N_{\mathcal{T}} = 0_{m\times m}$ and for $t = 1, ..., \mathcal{T}$ *-1* we use

$$\begin{aligned}
K_t &= TPZ'F_t^{-1} && (16)\\
L_t &= T - K_t Z_t && (17)\\
r_{t-1} &= Z_t'F_t^{-1}v_t + L_t' r_t && (18)\\
N_{t-1} &= Z_t'F_t^{-1}Z_t + L_t' N_t L_t && (19)\\
\hat{\alpha}_t &= a_t + P_t r_{t-1} && (20)\\
V_t &= P_t - P_t N_{t-1} P_t, && (21)
\end{aligned}$$

where we store the matrices $\{F_t, v_t, a_t, P_t\}_{t=1}^{\mathcal{T}}$ from one run of the Kalman filter. We use the appropriate element of the $\hat{\alpha}_t$ vector as the extracted factor and the corresponding diagonal element of V_t as its standard error to compute confidence bands.

Before we turn to estimation, we describe how the Kalman filter handles missing observations. If for a period t, all elements of the vector Y_t are missing, we skip updating and the recursion becomes

$$\begin{aligned}
a_{t+1} &= Ta_t && (22)\\
P_{t+1} &= TP_tT' + RQR. && (23)
\end{aligned}$$

If some (but not all) elements of Y_t are missing, we replace the observation equation with[8]

$$Y_t^* = Z_t^* \alpha_t + \Lambda X_t + \varepsilon_t^* \tag{24}$$

$$\varepsilon_t^* \sim N\left(0, H_t^*\right), \tag{25}$$

where Y_t^* are the elements of the Y_t vector that are observed. The two vectors are linked by $Y_t^* = W_t Y_t$ where W_t is a matrix that carries the appropriate rows or $I_{p\times p}$, $Z_t^* = W_t Z_t$, $\varepsilon_t^* = W_t \varepsilon_t$ and $H_t^* = W_t H_t W_t'$. The Kalman filter and smoother work exactly as described above replacing Y_t, Z_t and H with Y_t^*, Z_t^* and H_t^* for period t.

In calculating the log likelihood, if all elements of Y_t are missing, the contribution of period t to the likelihood is zero. When some elements of Y_t are observed, the contribution of period t will be $\left[p^* \log 2\pi + \left(\log |F_t^*| + v_t^{*\prime} F_t^{*-1} v_t^*\right)\right]$ where p^* is the number of observed variables and the other matrices and vectors are obtained using the Kalman filter recursions on the modified system with Y_t^*.

2.4 Estimation

2.4.1 Classical Implementation

The latent factor is stationarity in our framework. To impose that constraint on our estimates we use a result of Barndorff-Nielsen and Schou (1973), who show that under stationarity there is a one-to-one correspondence between the parameters of an AR(p) process and the first p partial autocorrelations. Hence we can parameterize the likelihood in terms of the relevant partial autocorrelations, which require searching only over the unit interval.

In our subsequent empirical analysis we use an $AR(3)$ process for the factor, which allows for one real root and two imaginary roots and hence a rich variety of dynamics. Denoting the AR(3) parameters by ρ_i and the partial autocorrelations by π_i, the Barndorff-Nielsen-Schou mapping between the two is

$$\rho_1 = \pi_1 - \pi_1\pi_2 - \pi_3\pi_2 \tag{26}$$

$$\rho_2 = \pi_2 - \pi_1\pi_3 + \pi_1\pi_2\pi_3 \tag{27}$$

$$\rho_3 = \pi_3. \tag{28}$$

[8]By construction, whenever there is an observation for a particular element of Y_t, there is a corresponding element of X_t.

We then optimize over $\pi_i \in [-1, 1]$.[9]

We also use two restrictions in our estimation. We guarantee the non-negativity of the variance terms in the diagonal elements of Q and H matrices by estimating natural logarithms of these elements. We also restrict the factor loading on some of the variables to have a certain sign (e.g. positive for GDP and negative for initial jobless claims) using the same transformation.

Once we obtain the log likelihood for a given set of parameters, we can proceed with estimation using standard methods. In particular, we use a quasi-Newton optimization routine with BFGS update of the inverse Hessian.

Searching for a global optimum in a parameter space with more than 30 dimensions is a challenging problem. It is not intractable, however, if the the iterations are initialized cleverly. To do so, we exploit knowledge gained from certain auxilliary regressions, as well as a series of pilot experiments, which we describe in section 3.

2.4.2 Bayesian Implementation

[To be completed]

3 Examples

We provide a number of examples to illustrate the details of our method before we proceed to analyze a full model in the next section. We first show the effectiveness of our method in a simulation exercise, and then we proceed to a "toy" example with real data which resembles a simplified version of our full model, followed by a more complete and serious example.

3.1 A Simple Simulation Example

We use this example to demonstrate how our methodology works. We use a simplified version of the setup in our full model to generate an artificial data set. Specifically, we assume that the true daily factor follows an AR(1) process and that 3 daily variables are linked to this daily factor and a linear trend term. For simplicity we do not use the PDL specification, higher order trend terms or lagged dependent variables. We generate 40 years' worth of daily data, which roughly corresponds to our actual data set. After we generate the daily data, we transform them to obtain the dataset that the economist observes. For y_t^1, which

[9] We use a hyperbolic tangent function to search over π_i, because for $y \in R$, $x = \tanh(y) \in [-1, 1]$.

is a daily financial variable, we eliminate the observations for the weekends. For y_t^2, which is a monthly stock variable we eliminate all the observations except for the last observation for each month. Finally, for y_t^3, which is a quarterly flow variable, all observations except for the last observation in the quarter are missing and the last observation of the quarter is simply the sum of all the daily observations. After obtaining the data set we estimate the state space system described in (7) where the system vectors and matrices are defined as follows:

$$Y_t = \begin{bmatrix} \tilde{y}_t^1 \\ \tilde{y}_t^2 \\ \tilde{y}_t^3 \end{bmatrix}, \quad \alpha_t = \begin{bmatrix} x_t \\ x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-q_{\max}+1} \\ x_{t-q_{\max}} \end{bmatrix}, \quad X_t = \begin{bmatrix} 1 \\ t \end{bmatrix}, \quad \varepsilon_t = \begin{bmatrix} \varepsilon_t^2 \\ \varepsilon_t^2 \\ \varepsilon_t^{*3} \end{bmatrix}, \quad v_t = \eta_t, \quad R = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{29}$$

$$Z = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 & 0 & 0 \\ \beta_2 & 0 & \cdots & 0 & 0 & 0 \\ \beta_3 & \beta_3 & \cdots & \beta_3 \text{ or } 0 & \beta_3 \text{ or } 0 & \beta_3 \text{ or } 0 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} c_1 & \delta_1 \\ c_2 & \delta_2 \\ c_3^* & \delta_3^* \end{bmatrix}, \quad T = \begin{bmatrix} \rho_1 & \rho_2 & \rho_3 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \varepsilon_t \\ v_t \end{bmatrix} \sim N\left(\begin{bmatrix} 0_{3\times 1} \\ 0 \end{bmatrix}, \begin{bmatrix} H & 0 \\ 0 & Q \end{bmatrix} \right), \quad H = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^{*2} \end{bmatrix}, \quad Q = 1$$

where $q_{\max}$ is the maximum number of days in a quarter. Even though in the notation we treat q, the number of days in a quarter (the counterpart of D_i in the discussion in the previous section) as a fixed number, in our implementation we make the necessary adjustments to take into account the exact number of days in a quarter.[10] All of the relevant matrices and vectors allow for the largest possible value, $q_{\max}$ and we adjust the matrices

[10]This number is either 90, 91 or 92, depending on the quarter and whether or not the year is a leap year.

Z, Λ and H every quarter in the following way.[11] At each quarter, if $q < q_{\max}$ the first q elements of the third row of Z are set to β_3 while the remaining elements are set equal to zero. Next, we use $D = q$ in the formulas derived in the appendix that map our original parameters c_3, δ_{31} to c_3^*, δ_{31}^* and substitute in Λ. Finally, the third diagonal element of H is set to $\sigma_3^{*2} = q\sigma_3^2$.[12]

To mimic our estimation procedure with real data we first estimate a smaller version of this model: we use only the first two variables. Once we have estimates from this model, we use the Kalman smoother to extract the factor. We then run the auxiliary regression

$$\tilde{y}_t^3 = \sum_{j=0}^{q-1} \left[a + d\left(t - j\right)\right] + b\left(x_t + x_{t-1} + ... + x_{t-q}\right) + e_t \tag{30}$$

using OLS and use the estimates for a, b, d and $var\left(e_t\right)/q_{\max}$ as the starting values for the estimation of the full model for c_3, β_3, δ_3 and σ_3^2 while for all the other parameters in the full model common to the smaller model we use the estimated values from the smaller model.[13]

In Table 1 we report the estimation results from the two stages along with the true parameters that we used to generate the data. The first column reports the true parameters used in the simulation and the second column reports the in-sample values for these parameters, obtained by running OLS using the simulated time series for x_t, y_t^1, y_t^2, and y_t^3 which do not match the true values due to sampling error given our "small" sample. The third column reports the estimates from the smaller system, the next column reports the results from the auxiliary regression (30) and the last column reports the estimates from the full system, using the parameters from the previous two columns as starting values for the optimization routine. We also report the values of the log likelihood from each estimation. Comparing the estimates from stage 2 to the true parameters, we see that with the possible exception of c_1 and c_2 the estimates are almost identical to the true values, up to a rounding error and the estimates of c_1 and c_2 are fairly close to their true counterparts. It is also interesting to point out that the $\log L$ for Stage 2 is very close to the sum of the $\log L$ of Stage 1 and the auxiliary model. We will use this fact as an informal diagnostics tool when

[11] The third rows of Z, Λ and H are only relevant when $\tilde{y}_t^3$ is observed. For all other days, the contents of the third rows of these matrices does not affect any calculations. When there is an observation for $\tilde{y}_t^3$ we look at the number of days in that particular quarter, q, and make the adjustments.

[12] These all follow from the discussion in the previous section. The quarterly flow variable requires summing the factors over the quarter and setting some of the elements of the third row of Z would make sure we sum only the relevant factors. The adjustment of the trend coefficients should be obvious. Finally, since ε_t^{*3} is the sum of q iid normal innovations each with variance σ_3^2, its variance is $q\sigma_3^2$.

[13] In this regression we adjust q according to the actual number of days in the quarter.

estimating our full model.

Perhaps more important than the parameter estimates is the ability of our methodology to extract a factor which is close to the true factor we used to generate the data. Table 2 reports the correlations of the true factor with three smoothed factors using our methodology: using the in-sample values for the parameters (with no estimation), from Stage 1 and from Stage 2. All correlations are greater than 0.96, which shows that under these fairly realistic conditions our methodology is able to extract a factor which is very close to the true one.

To illustrate the performance of our methodology we also plot the observed, smoothed and true versions for the first two signals (daily financial variable and monthly stock variable) over a 6-month period. In the first panel, the observed and the true signals are identical except for the weekends, and the smoothed signal tries to fill in the missing values in the observed signal by using the information from other variables. In the second panel, the observed signal is represented by blue dots which are the end-of-month-values of the true signal. Our smoothed signal tries to fill in the other values and performs quite well. Overall, this example shows that our methodology is well-suited to extract the factor in an environment with missing data and/or time aggregation issues.

3.2 A Simple Empirical Example

Our first example with real data uses daily observations on the slope of the yield curve (the term premium) and monthly observations on payroll employment.[14] The state variable x_t follows an AR(3) process and we assume an AR(3) structure for both observed variables at their observation frequency. For monthly employment, this means the value of employment in the previous 3 months is an element of the X_t vector and we denote these by $\tilde{y}^2_{t-M}$, $\tilde{y}^2_{t-2M}$ and $\tilde{y}^2_{t-3M}$.[15] For the term premium, on the other hand, we choose to model this structure by assuming an AR(3) process for the measurement equation innovation. If we had no missing observations for the term premium, either method would yield identical results. We choose to follow this route because of the missing term premium observations due to non-business days. If we used the lagged term premium as an element of X_t this would mean we only have 2 valid observations for each week and it would make the analysis less reliable.[16] As a

[14] See Section 4.1 for details on the data.

[15] Once again, the notation in the paper assumes M is constant over time but in the implementation we adjust M according to the number of days in a month.

[16] Alternatively we could have used AR(3) measurement errors for all variables. But this persistence in the daily frequency would essentially disappear when we aggregate the variables to the monthly or quarterly frequency.

result, the system we use for this example can be summarized as follows

$$Y_t = \begin{bmatrix} \tilde{y}_t^1 \\ \tilde{y}_t^2 \end{bmatrix}, \quad \alpha_t = \begin{bmatrix} x_t \\ x_{t-1} \\ x_{t-2} \\ \varepsilon_t^1 \\ \varepsilon_{t-1}^1 \\ \varepsilon_{t-2}^1 \end{bmatrix}, \quad X_t = \begin{bmatrix} 1 \\ t \\ t^2 \\ t^3 \\ \tilde{y}_{t-M}^2 \\ \tilde{y}_{t-2M}^2 \\ \tilde{y}_{t-3M}^2 \end{bmatrix}, \quad \varepsilon_t = \begin{bmatrix} 0 \\ \varepsilon_t^2 \end{bmatrix}, \quad v_t = \begin{bmatrix} \eta_t \\ v_t^1 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$Z = \begin{bmatrix} \beta_1 & \beta_2 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}', \quad \Lambda = \begin{bmatrix} c_1 & c_2 \\ \delta_{11} & \delta_{21} \\ \delta_{12} & \delta_{22} \\ \delta_{13} & \delta_{23} \\ 0 & \gamma_{21} \\ 0 & \gamma_{22} \\ 0 & \gamma_{23} \end{bmatrix}', \quad T = \begin{bmatrix} \rho_1 & \rho_2 & \rho_3 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{31}$$

$$\begin{bmatrix} \varepsilon_t \\ v_t \end{bmatrix} \sim N\left(\begin{bmatrix} 0_{2\times 1} \\ 0_{2\times 1} \end{bmatrix}, \begin{bmatrix} H & 0 \\ 0 & Q \end{bmatrix} \right), \quad H = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}$$

where the matrices and vectors correspond to the system in Section 2.2 and we have $p = 2$, $k = 7$, $m = 6$ and $r = 2$. When estimating this system, we restrict β_2 to be positive to impose a positive relationship between the factor and employment.

The parameter estimates for this system along with other statistics are provided in the Appendix. We will turn to the smooth factor from this example when we complete the analysis of the full model below.

4 A Four-Variable Model

4.1 Data

Our analysis covers the period from April 1, 1962 through February 20, 2007, which is over 45 years of daily data. Since assuming economic activity stops over the weekends is not realistic, we use a 7-day week instead of using only business days. We use 4 variables in our analysis. Below we list these variables along with how we handle the missing data / time

aggregation issues.

- *Yield curve term premium defined as the difference between the yield of the 10-year and the 3-month Treasury yields.* [TERM] This is a daily variable.
- *Average weekly initial claims for unemployment insurance.* [IJC] This is a weekly flow variable covering the 7-day period from Sunday through Saturday. The value for Saturdays is the sum of the daily values for the previous 7-days.
- *Employees on nonagricultural payrolls.* [EMP] This is a monthly stock variable, observed on the last day of the month.
- *Real GDP.* [GDP] This is a quarterly flow variable. The value for the last day of the quarter is the sum of the daily values for all the days in the quarter.

For numerical stability we adjust the units of some of our observed variables. (e.g. we divide EMP by 10,000 and IJC by 1,000)

4.2 Model

Ordering the observed variables in decreasing frequency, The matrices that define the full model are given by

$$Y_t = \begin{bmatrix} \tilde{y}_t^1 \\ \tilde{y}_t^2 \\ \tilde{y}_t^3 \\ \tilde{y}_t^4 \end{bmatrix}, \quad \alpha_t = \begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-\bar{q}-1} \\ x_{t-\bar{q}} \\ \varepsilon_t^1 \\ \varepsilon_{t-1}^1 \\ \varepsilon_{t-2}^1 \end{bmatrix}, \quad X_t = \begin{bmatrix} 1 \\ t \\ t^2 \\ t^3 \\ \tilde{y}_{t-W}^2 \\ \tilde{y}_{t-2W}^2 \\ \tilde{y}_{t-3W}^2 \\ \tilde{y}_{t-M}^3 \\ \tilde{y}_{t-2M}^3 \\ \tilde{y}_{t-3M}^3 \\ \tilde{y}_{t-q}^4 \\ \tilde{y}_{t-2q}^4 \\ \tilde{y}_{t-3q}^4 \end{bmatrix}, \quad \varepsilon_t = \begin{bmatrix} 0 \\ \varepsilon_t^2 \\ \varepsilon_t^3 \\ \varepsilon_t^4 \end{bmatrix}, \quad v_t = \begin{bmatrix} \eta_t \\ v_t^1 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$Z = \begin{bmatrix} \beta_1^0 & \beta_2 & \beta_3 & \beta_4 \\ \beta_1^1 & \beta_2 & 0 & \beta_4 \\ \vdots & \vdots & \vdots & \vdots \\ \beta_1^6 & \beta_2 & 0 & \beta_4 \\ \beta_1^7 & 0 & 0 & \beta_4 \\ \vdots & \vdots & \vdots & \vdots \\ \beta_1^{\bar{q}-1} & 0 & 0 & \beta_4 \\ \beta_1^{\bar{q}} & 0 & 0 & \beta_4 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}', \quad \Lambda = \begin{bmatrix} c_1 & c_2^* & c_3 & c_4^* \\ \delta_{11} & \delta_{21}^* & \delta_{31} & \delta_{41}^* \\ \delta_{12} & \delta_{22}^* & \delta_{32} & \delta_{42}^* \\ \delta_{13} & \delta_{23}^* & \delta_{33} & \delta_{43}^* \\ 0 & \gamma_{21} & 0 & 0 \\ 0 & \gamma_{22} & 0 & 0 \\ 0 & \gamma_{23} & 0 & 0 \\ 0 & 0 & \gamma_{31} & 0 \\ 0 & 0 & \gamma_{32} & 0 \\ 0 & 0 & \gamma_{33} & 0 \\ 0 & 0 & 0 & \gamma_{41} \\ 0 & 0 & 0 & \gamma_{42} \\ 0 & 0 & 0 & \gamma_{43} \end{bmatrix}' \tag{32}$$

$$T = \begin{bmatrix} \rho_1 & \rho_2 & \rho_3 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \varepsilon_t \\ v_t \end{bmatrix} \sim N\left(\begin{bmatrix} 0_{4\times 1} \\ 0_{2\times 1} \end{bmatrix}, \begin{bmatrix} H & 0 \\ 0 & Q \end{bmatrix} \right), \quad H = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \sigma_2^{*2} & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_{*4}^2 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}$$

where the matrices and vectors correspond to the system in Section 2.2 and we have $p = 4$, $k = 13$, $m = 95$ and $r = 2$. W denotes the number of days in a week, M denotes the number of days in a month and q denotes the number of days in a quarter. While $W = 7$, M and q vary according to the specific month and quarter. As we did in the artificial data example, we use the transformation given in the appendix to convert the coefficients with $*$ to those without.

We use the current and 91 lags of the factor in our state space system since the maximum of days possible in a quarter is 92, which we denote by $\bar{q}$.[17] In every quarter, we adjust the number of non-zero elements in the fourth row of the Z matrix to reflect the number of days in that quarter. When estimating this system, we restrict β_3 and β_4 to be positive and β_2 to be negative to reflect our expectation of the relationship between these variables and the common factor.[18]

4.3 Classical Results

We use the classical methods described above to this model. It is worth emphasizing the size of this model. We have 16,397 daily observations, 95 state variables and 42 coefficients. Using a fairly efficiently programmed Kalman filter routine in MATLAB, one evaluation of the log-likelihood takes about 25 seconds. As such, one iteration (including the calculation of the Jacobian) takes a minimum of 18 minutes. Clearly, it is very costly to look over an "irrelevant" part of the parameter space as it may take the estimation routine many hours or days to find the "right" path, if at all. To tackle this problem, we follow the algorithm outlined earlier: We start by a smaller system, one that has only TERM and EMP. Once we estimate this system we get the smoothed factor and estimate the auxiliary regression for GDP. Using the estimated values from the smaller system and the auxiliary regression as the starting guesses, we estimate the system with GDP. We repeat this for IJC.

Here we focus on the factor and its properties.[19] Figure 2 plots the smoothed factor from the estimation along with 95% error bands, with the NBER recession dates superimposed. Since the NBER provides only months, we assume recessions start on the first day of the month and end on the last day of the month. There a few important observations. First, the smoothed factor declines sharply around the recessions dates announced by the NBER. The beginning of recessions and the downturn of the smoothed factor does not always coincide however, the factor shows the same sharp decline pattern at the start of each of the 6 recessions in the sample. There can also be slight mismatch due to the monthly structure of the NBER's recession timing. Second, recoveries show different patterns. For the recessions in 1974, 1980 and 1982 the recoveries are almost as sharp as the declines. For the three

[17] If there are Q days in a quarter, on the last day of the quarter, we need the current and the $Q-1$ lags of the factor for the measurement equation of GDP.

[18] In our experience with smaller systems, when we do not impose a sign restriction the estimation may yield a factor which is negatively correlated with GDP. Imposing the sign restriction reverses the correlation with virtually no change in the likelihood.

[19] The parameter estimates for the full system are reported in the appendix.

remaining recessions, as well as the 1961 recession which is just before our sample starts, the recoveries are slower, especially so for the 1990 recession as is well-known. Fourth, it seems that the recovery from the last recession in our sample displays different characteristics from the previous one. The latter displays a slow but consistent recovery while the former stays low for about 2 years after the official end of the recession, even going down after a small recovery. It seems by the end of our sample February 2006, the economy has recovered from the recession, following a sharp increase in the factor in 2004. Finally, there seems to be few, if any, "false positives" where our factor shows patterns similar to recessions in a period which is not a recession. Overall, we conclude that our smoothed factor is able to follow the business cycles of the US in our sample very well.

Figure 3 plots the smooth factors from the term premium / employment example and the full model. The two factors have a correlation of 0.86 and while they agree on the turning points for the most part, they show differences regarding the extent of recessions for especially the last two recessions. There are two main differences in the models for these two factors. First, the full model uses information from GDP. Second, the full model uses the PDL structure for the term premium and allows for a richer interaction between the yield curve and the aggregate economic activity. To see the effect of the latter, we plot the estimated PDL coefficients for the term premium in Figure 4. We see a strong negative relationship between the term premium at time t and the factor at time $t - s$ for about $s = 20$ days.

4.4 Bayesian Results

[To be completed.]

5 Summary and Concluding Remarks

We have constructed a framework for measuring macroeconomic activity in real time, using a variety of stock and flow data observed at mixed frequencies, including ultra-high frequencies. Specifically, we have proposed a dynamic factor model that permits exactly optimal extraction of the latent state of macroeconomic activity, and we have illustrated it both in simulation environments and in a sequence of progressively richer empirical examples.

We look forward to a variety of variations and extensions of our basic theme, including but not limited to:

(1) Incorporation of indicators beyond macroeconomic and financial data. In particular, it will be of interest to attempt inclusion of qualitative information such as headline news.

(2) Construction of a real time composite leading index (CLI). Thus far we have focused only on construction of a composite *coincident index* (CCI), which is the more fundamental problem, because a CLI is simply a forecast of a CCI. Explicit construction of a leading index will nevertheless be of interest.

(3) Allowance for nonlinear regime-switching dynamics. The linear methods used in this paper provide only a partial (linear) statistical distillation of the rich business cycle literature. A more complete approach would incorporate the insight that expansions and contractions may be probabilistically different regimes, separated by the "turning points" corresponding to peaks and troughs, as emphasized for many decades in the business cycle literature and rigorously embodied Hamilton's (1989) Markov-switching model. Diebold and Rudebusch (1996) and Kim and Nelson (1998) show that the linear and nonlinear traditions can be naturally joined via dynamic factor modeling with a regime-switching factor. Such an approach could be productively implemented in the present context, particularly if interest centers on turning points, which are intrinsically well-defined only in regime-switching environments.

(4) Comparative assessment of experiences and results from "small data" approaches, such as ours, vs. "big data" approaches. Although much professional attention has recently turned to big data approaches, as for example in Forni, Hallin, Lippi and Reichlin (2000) and Stock and Watson (2002), recent theoretical work by Boivin and Ng (2006) shows that bigger is not necessarily better. The matter is ultimately empirical, requiring detailed comparative assessment. It would be of great interest, for example, to compare results from our approach to those from the Altissimo *et al.*(2002) EuroCOIN approach, for the same economy and time period. Such comparisons are very difficult, of course, because the "true" state of the economy is never known, even ex post.

References

[1] Abeysinghe, T. (2000), "Modeling Variables of Different Frequencies," International Journal of Forecasting, 16, 117-119.

[2] Altissimo, F., Bassanetti, A., Cristadoro, R., Forni, M., Hallin, M., Lippi, M., Reichlin, L. and Veronese, G. (2001), "Eurocoin: A Real Time Coincident Indicator of the Euro Area Business Cycle," CEPR Discussion Paper No. 3108.

[3] Boivin, J. and Ng, S. (2006), "Are More Data Always Better for Factor Analysis," Journal of Econometrics, 127, 169-194.

[4] Burns, A.F. and Mitchell, W.C. (1946), Measuring Business Cycles, New York, NBER.

[5] Diebold, F.X. (2003), "'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting" (Discussion of Reichlin and Watson papers), in M. Dewatripont, L.P. Hansen and S. Turnovsky (Eds.), Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society. Cambridge: Cambridge University Press, 115 122.

[6] Diebold, F.X. and Rudebusch, G. (1996), "Measuring Business Cycles: A Modern Perspective," Review of Economics and Statistics, 78, 67 77.

[7] Durbin and Koopman (2001)

[8] Evans, M.D.D. (2005), "Where Are We Now?: Real Time Estimates of the Macro Economy," The International Journal of Central Banking, September.

[9] Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000), "The Generalized Factor Model: Identification and Estimation," Review of Economics and Statistics, 82, 540-554.

[10] Geweke, J.F. (1977), "The Dynamic Factor Analysis of Economic Timeseries Models," in D. Aigner and A. Goldberger (eds.), Latent Variables in Socio economic Models, North Holland, 1977, pp. 365 383.

[11] Ghysels, E., Santa-Clara, P. and Valkanov, R.(2004), "The MIDAS Touch: Mixed Data Sampling Regression Models," Manuscript, University of North Carolina.

[12] Hall, R.E., et al. (2003), "he NBER's Recession Dating Procedure," http://www.nber.org/cycles/recessions.html

[13] Hamilton, J.D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," Econometrica, 57, 357 384.

[14] Kim, C.-J. and Nelson, C.R. (1998), State Space Models with Regime Switching: Classical and Gibbs Sampling Approaches with Applications. Cambridge, Mass.: MIT Press.

[15] Liu, H. and Hall, S.G. (2001), "Creating High-frequency National Accounts with State-Space Modelling: A Monte Carlo Experiment,"Journal of Forecasting, 20, 441-449.

[16] Lucas, R.E. (1977), "Understanding Business Cycles," Carnegie Rochester Conference Series on Public Policy, 5, 7 29.

[17] Mariano, R.S. and Murasawa, Y. (2003), "A New Coincident Index of Business Cycles Based on Monthly and Quarterly Series," Journal of Applied Econometrics, 18, 427 443.

[18] McGuckin, R.H., Ozyildirim, A. and Zarnowitz, V. (2003), "A More Timely and Useful Index of Leading Indicators," Manuscript, Conference Board.

[19] Proietti, T. and Moauro, F. (2006), "Dynamic Factor Analysis with Non Linear Temporal Aggregation Constraints," Applied Statistics, 55, 281-300.

[20] Sargent, T.J. and Sims, C.A. (1977), "Business Cycle Modeling Without Pretending to Have Too Much A Priori Economic Theory," in C. Sims (ed.), New Methods in Business Research. Minneapolis: Federal Reserve Bank of Minneapolis.

[21] Shen, C.-H. (1996), "Forecasting Macroeconomic Variables Using Data of Different Periodicities," International Journal of Forecasting, 12, 269-282.

[22] Stock, J.H. and Watson, M.W. (1989), "New Indexes of Coincident and Leading Economic Indicators," NBER Macro Annual, Volume 4. Cambridge, Mass.: MIT Press.

[23] Stock, J.H. and Watson, M.W. (1991), "A Probability Model of the Coincident Economic Indicators." In K. Lahiri and G. Moore (eds.), Leading Economic Indicators: New Approaches and Forecasting Records. Cambridge: Cambridge University Press, 63 89.

[24] Stock, J.H. and Watson, M.W. (2002), "Macroeconomic Forecasting Using Diffusion Indexes," Journal of Business and Economic Statistics, 20, 147-162.

Appendix

A Mapping for the Constant and the Coefficients of Trend

Here we establish the mapping between two sets of parameters. On the one hand, we have

$$\sum_{j=0}^{D-1}\left[c+\delta_1\left(\frac{t-j}{1000}\right)+\delta_2\left(\frac{t-j}{1000}\right)^2+\delta_3\left(\frac{t-j}{1000}\right)^3\right] \tag{33}$$

and on the other hand we have

$$c^*+\delta_1^*\left(\frac{t}{1000}\right)+\delta_2^*\left(\frac{t}{1000}\right)^2+\delta_3^*\left(\frac{t}{1000}\right)^3 \tag{34}$$

We want to establish the mapping between $(c,\delta_1,\delta_2,\delta_3)$ and $(c^*,\delta_1^*,\delta_2^*,\delta_3^*)$.

$$\sum_{j=0}^{D-1}\left[c+\delta_1\left(\frac{t-j}{1000}\right)+\delta_2\left(\frac{t-j}{1000}\right)^2+\delta_3\left(\frac{t-j}{1000}\right)^3\right] \tag{35}$$

$$= \sum_{j=0}^{D-1}c+\delta_1\sum_{j=0}^{D-1}\left(\frac{t}{1000}-\frac{j}{1000}\right)+\delta_2\sum_{j=0}^{D-1}\left(\frac{t}{1000}-\frac{j}{1000}\right)^2+\delta_3\sum_{j=0}^{D-1}\left(\frac{t}{1000}-\frac{j}{1000}\right)^3 \tag{36}$$

$$= Dc+\delta_1\sum_{j=0}^{D-1}\left(\frac{t}{1000}\right)-\delta_1\sum_{j=0}^{D-1}\left(\frac{j}{1000}\right) \tag{37}$$

$$+\delta_2\sum_{j=0}^{D-1}\left(\frac{t}{1000}\right)^2+\delta_2\sum_{j=0}^{D-1}\left(\frac{j}{1000}\right)^2-2\delta_2\sum_{j=0}^{D-1}\frac{tj}{1000^2} \tag{38}$$

$$+\delta_3\sum_{j=0}^{D-1}\left(\frac{t}{1000}\right)^3-\delta_3\sum_{j=0}^{D-1}\left(\frac{j}{1000}\right)^3-3\delta_3\sum_{j=0}^{D-1}\left(\frac{t}{1000}\right)^2\left(\frac{j}{1000}\right)+3\delta_3\sum_{j=0}^{D-1}\left(\frac{t}{1000}\right)\left(\frac{j}{1000}\right)^2 \tag{39}$$

$$
\begin{aligned}
&= Dc - \delta_1 \sum_{j=0}^{D-1} \left(\frac{j}{1000}\right) + \delta_2 \sum_{j=0}^{D-1} \left(\frac{j}{1000}\right)^2 - \delta_3 \sum_{j=0}^{D-1} \left(\frac{j}{1000}\right)^3 && (40)\\
&+ \frac{t}{1000}\left[D\delta_1 - 2\delta_2 \sum_{j=0}^{D-1} \frac{j}{1000} + 3\delta_3 \sum_{j=0}^{D-1} \left(\frac{j}{1000}\right)^2\right] && (41)\\
&+ \left(\frac{t}{1000}\right)^2 \left[D\delta_2 - 3\delta_3 \sum_{j=0}^{D-1} \left(\frac{j}{1000}\right)\right] && (42)\\
&+ \left(\frac{t}{1000}\right)^3 (D\delta_3) && (43)
\end{aligned}
$$

Now, note the formulas

$$
\begin{aligned}
\sum_{j=0}^{D-1} j &= \frac{D\left(D-1\right)}{2} && (44)\\
\sum_{j=0}^{D-1} j^2 &= \frac{D\left(D-1\right)\left[2\left(D-1\right)+1\right]}{6} = \frac{D\left(D-1\right)\left(2D-1\right)}{6} && (45)\\
\sum_{j=0}^{D-1} j^3 &= \left[\frac{D\left(D-1\right)}{2}\right]^2 && (46)
\end{aligned}
$$

So we get

$$
\begin{aligned}
c^* &= Dc - \frac{\delta_1 D\left(D-1\right)}{2000} + \frac{\delta_2 D\left(D-1\right)\left(2D-1\right)}{6\times 10^6} - \frac{\delta_3\left[D\left(D-1\right)\right]^2}{4\times 10^9} && (47)\\
\delta_1^* &= D\delta_1 - \frac{\delta_2 D\left(D-1\right)}{1000} + \frac{\delta_3 D\left(D-1\right)\left(2D-1\right)}{2\times 10^6} && (48)\\
\delta_2^* &= D\delta_2 - \frac{3\delta_3^* D\left(D-1\right)}{2000} && (49)\\
\delta_3^* &= D\delta_3 && (50)
\end{aligned}
$$

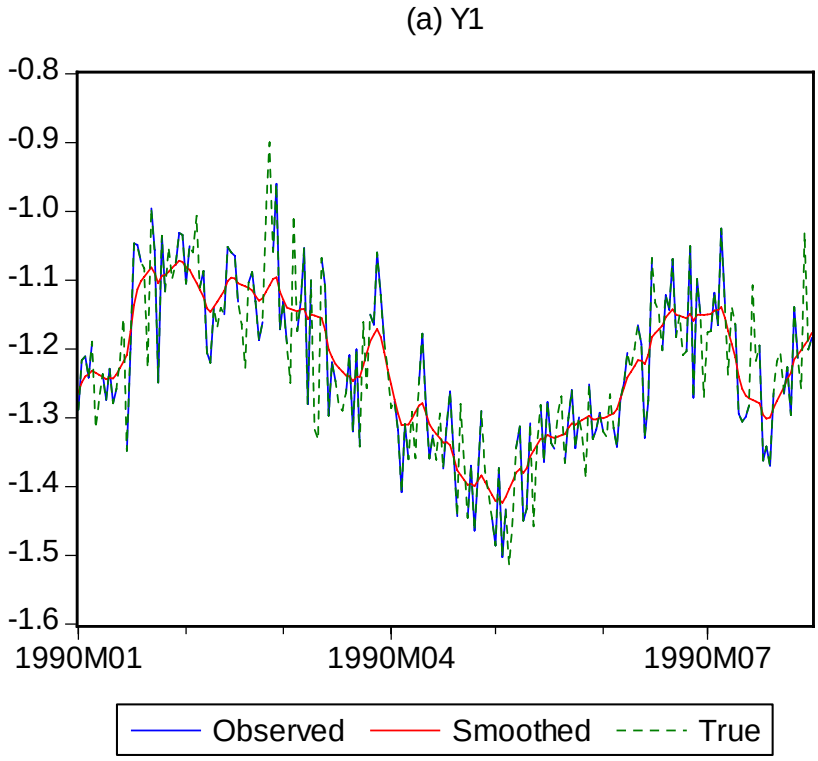
Table 1 - Parameter Values for the Example with Artificial Data

Parameter	True Values	In-Sample Values	Estimate Stage 1	Auxilary	Stage 2
c_1	0.9	0.9012	1.0262 (0.057)	–	1.0474 ($xxxx$)
c_2	0.4	0.4002	0.3963 (0.002)	–	0.3956 ($xxxx$)
c_3	-0.003	-0.0029	–	-0.0071 (0.000)	-0.0078 ($xxxx$)
β_1	-0.03	-0.0300	-0.0297 (0.000)	–	-0.0297 ($xxxx$)
β_2	0.001	0.0010	0.0010 (0.000)	–	0.0009 ($xxxx$)
β_3	0.001	0.0010	–	0.0009 (0.000)	0.0010 ($xxxx$)
δ_1	-0.2	-0.2000	-0.2102 (0.005)	–	-0.2120 ($xxxx$)
δ_2	0.03	0.0300	0.0304 (0.000)	–	0.0291 ($xxxx$)
δ_3	0.02	0.0200	–	0.0203 (0.000)	0.0204 ($xxxx$)
ρ	0.99	0.9902	0.9895 (0.041)	–	0.9893 ($xxxx$)
σ_1^2	0.005	0.0051	0.0051 (0.000)	–	0.0051 ($xxxx$)
σ_2^2	0.0001	0.0001	0.0001 (0.000)	–	0.0001 ($xxxx$)
σ_3^2	0.00001	0.00001	-	0.00002 (0.000)	0.00002 ($xxxx$)
$logL$		12122.56	11841.78	284.93	12119.59

Table 2 - Correlations of Various Estimates of the Factor with the True Factor for the Example with Artificial Data

No Estimation (in-sample)	Stage 1	Stage 2
0.9860	0.9645	0.9634

Figure 1 – Signals from the Artificial Data Example



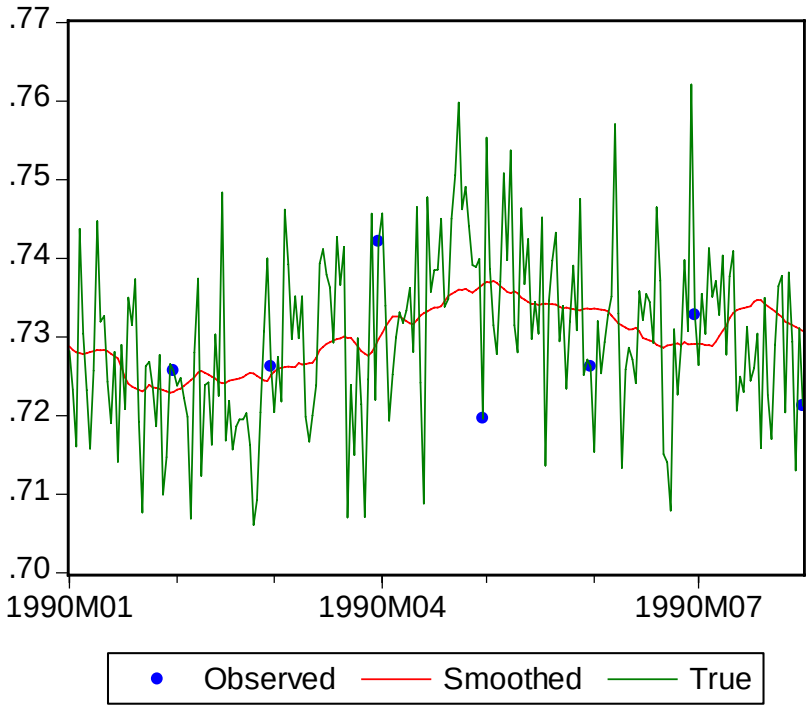


Figure 2 - Smoothed Factor from the Full Model

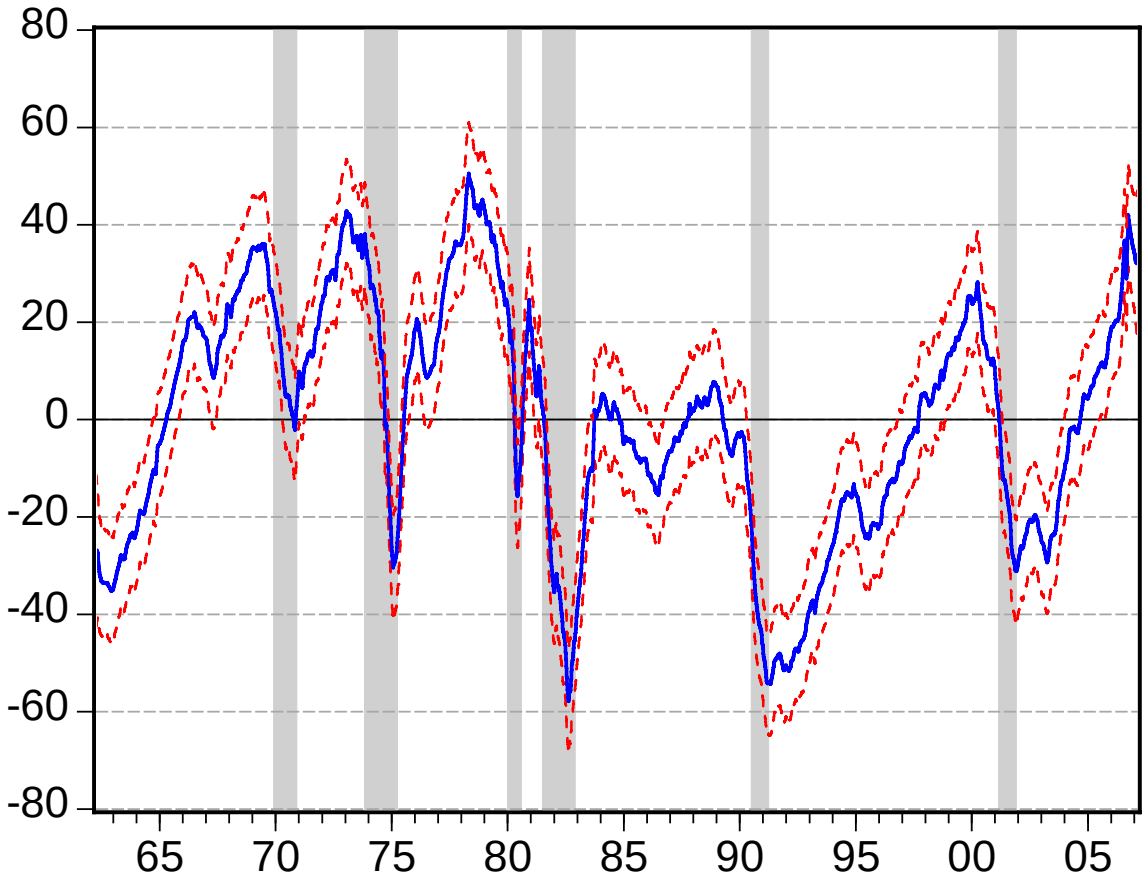


Figure 3 –Smoothed Factors

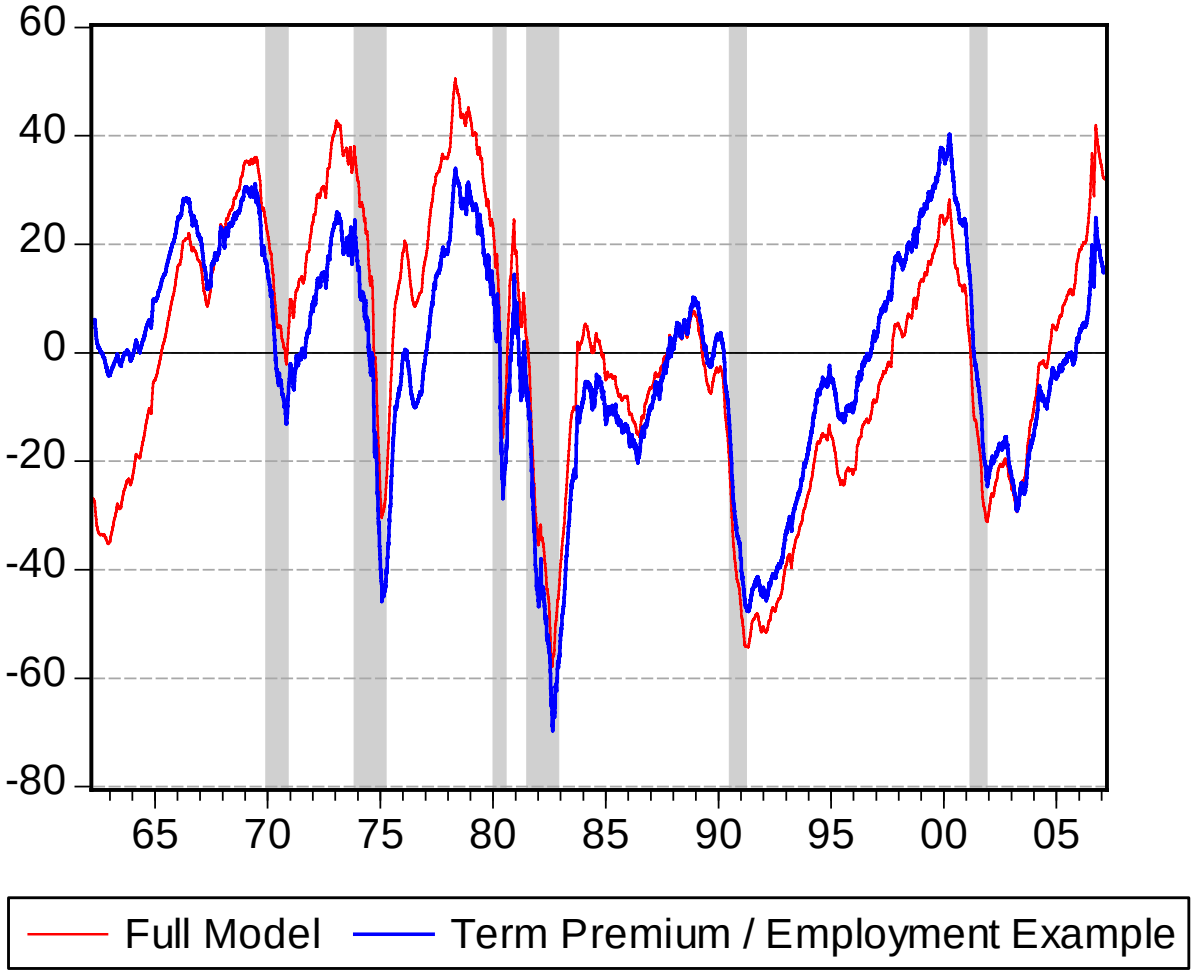


Figure 4 – Estimated Coefficients of the Polynomial Distributed Lag for Term Premium