Diffusion Index Model Specification and Estimation Using Mixed Frequency Datasets *

Kihwan Kim and Norman R. Swanson

Rutgers University

July 2012

Abstract

In this chapter, we discuss the use of mixed frequency models and diffusion index approximation methods in the context of prediction. In particular, select recent specification and estimation methods are outlined, and an empirical illustration is provided wherein U.S. unemployment forecasts are constructed using both classical principal components based diffusion indexes as well as using a combination of diffusion indexes and factors formed using small mixed frequency datasets. Preliminary evidence that mixed frequency based forecasting models yield improvements over standard fixed frequency models is presented.

JEL classification: C22, C51.

Keywords: forecasting, diffusion index, mixed frequency, recursive estimation, Kalman filter.

---

* Kihwan Kim and Norman R. Swanson, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA (kikim@econ.rutgers.edu and nswanson@econ.rutgers.edu). The authors would like to the editors, Jun Ma and Mark Wohar for inviting this contribution, and for providing useful comments. We would also like to thank Nii Ayi Armah, Valentina Corradi and Hyun Hak Kim for providing comments on earlier versions of this work.

1 Introduction

Economic time series datasets containing variables measured at varying frequencies have recently been increased usage amongst macroeconometricians. Two key approaches to specification and estimation of models incorporating variables of mixed frequency include the so-called MIxed DAta Sampling (MIDAS) regression approach, as discussed in Ghysels, Santa-Clara and Valkanov (2006) and Ghysels, Sinko and Valkanov (2006), and the references cited therein; and methods based on the classical state space representation proposed by Mariano and Murasawa (2003), which is refined and implemented in Aruoba, Diebold and Scotti (2009) and Aruoba and Diebold (2010). One interesting use for such models involves the estimation of factors that are subsequently used for constructing measures of "current economic activity"or for forecasting. An alternative to extracting common factors from mixed frequency datasets is to extract common factors (often called diffusion indexes) from largescale datasets, wherein all variables are measured at the same frequency, is discussed in Stock and Watson (2002a, 2006). The idea here is to extract a small number of "common" factors assumed to drive the dynamics associated with different policy-relevant and key forecasting variables. For example, applied practitioners, after estimating factor models, can subsequently use "key" diffusion indexes in the specification and estimation of forecasting models. Indeed, these sorts of "factor augmented forecasting models" have been found in the literature to yield predictions that often outperform those based on the specification of standard econometric models that do not include factors (see e.g. Armah and Swanson 2010, 2011, Kim and Swanson 2011, Stock and Watson 2002a,b, 2005, 2006, and the references cited therein). However, all of the above papers based on common factor methods focus on estimation, specification, and forecasting using datasets where all variables are of a single frequency. Given that the mixed frequency specification and estimation methods that are discussed above (and also discussed in detail in the sequel) allow for the convenient construction of diffusion indexes (i.e., factors) formed using variables of multiple different frequencies, a natural question is whether the combination of diffusion indexes based on both approaches yields improved prediction models. In this paper we review the extant literature in this area, and discuss simple approaches for addressing this question.

In order to illustrate the ideas discussed in this chapter, we also empirically examine a largescale dataset

and a small mixed frequency dataset in order to construct diffusion indexes to be used for forecasting U.S. unemployment.

In addition to the authors mentioned above, a number of researchers have recently made important contributions to the study of both dynamic and static common factor models specified with variables characterized by a common data measurement frequency. In these contexts, diffusion indexes are estimated using a variety of estimation techniques ranging from maximum likelihood to the Kalman filter. Key papers include Bai (2003), Bai and Ng (2002, 2006, 2010), Forni, Hallin, Lippi and Reichlin (2000, 2005), Hallin and Liska (2007), Onatski (2009, 2010) and Stock and Watson (2002b). Additionally, the properties of estimators based on generalized least squares are also discussed in Breitung and Tenhofen (2011), Doz and Reichlin (2011a) and Jungbacker and Koopman (2008). Doz and Reichlin (2011b) suggest a two step estimator by combining principal component and maximum likelihood methods. In order to evaluate the empirical usefulness of diffusion indexes in empirical applications, Stock and Watson (2009) examine diffusion index stability in regression contexts. Armah and Swanson (2010, 2011), Kim and Swanson (2011), and Stock and Watson (2002a, 2006) evaluate the usefulness of factor models in forecasting contexts, and Bernanke and Boivin (2003) use diffusion indexes to extract information useful for monetary policy evaluation.

As mentioned above, econometric researchers have recently been refining and further developing methods useful for extraction of common factors in mixed frequency datasets, with an eye to forecasting, nowcasting, and the use of so-called real-time data, whereby multiple revisions for each calendar dated observation are simultaneously modelled. , especially for nowcasting and forecasting. The MIxed DAta Sampling (MIDAS) regression approach (see e.g., Ghysels, Santa-Clara and Valkanov (2006) and Ghysels, Sinko and Valcanov (2006)), offers a complete methodology for estimation and inference using mixed frequency data. In earlier research, Mariano and Murasawa (2003) specify and estimate state space models in the same context. Recently, more general assumptions on factor dynamics (such as specification of generic ARMA processes) have been extensively examined by Mariano and Murasawa (2010), and Markov switching assumptions have been implemented in models discussed in Camacho, Pérez-Quirós and Poncela (2012). As discussed above, key recent papers include those by Aruoba, Diebold and Scotti (2009, henceforth ADS), Aruoba and Diebold (2010). Unlike contexts in which principal components are

extracted from largescale datasets, ADS (2009) assume that the latent process underlying their so-called "business conditions" index follows a simple process, such as an AR(1) process. In this context, ADS (2009) show that the business condition index constructed using a small but mixed frequency dataset mimics market fluctuations particularly well, especially during recession periods as announced by NBER.

To summarize, in this chapter we discussed fixed frequency and mixed frequency modeling, and present the results of a small empirical illustration in which U.S. unemployment is modeled using each approach, and using a combination of the two approaches. Interestingly, simple combination approaches, wherein mixed frequency diffusion indexes are combined with fixed frequency indexes, yield the mean square forecast error "best" predictions. The rest of the chapter is organized as follows. In section 2, we present our two dynamic factor modelling frameworks. In Section 3, we outline the empirical methodology used in our empirical illustration. Section 4 gathers the results of our empirical analysis, and concluding remarks are in Section 5.

2 The Modelling Framework

In this section, we recap a small subset of the factor modelling approaches discussed in a number of key papers, including Stock and Watson (1999, 2002a,b), Connor and Korajczyk (1986, 1988), and Forni, Hallin, Lippi and Reichlin (2000, 2005). The first is the dynamic factor modeling approach wherein principal components is used to estimate latent factors. These factors are called diffusion indexes in Stock and Watson (2002a). Thereafter, we discuss a mixed frequency dynamic factor model, which is estimated by maximum likelihood estimation in the spirit of ADS (2009) and Aruoba and Diebold (2010). For further review of dynamic factor models, see, for example, Armah and Swanson (2010) and Stock and Watson (2006, 2011).

2.1 Dynamic Factor Model

Following Stock and Watson (2006), suppose X_t has a dynamic factor model (henceforth DFM) representation with q common dynamic factors, f_t.

$$X_{it} = \lambda_i(L)' f_t + e_{it}, \tag{1}$$

for i = 1,2,...,N, and t = 1,2,...,T, where X_{it} is a single datum, f_t is the $q \times 1$ vector of unobserved factors, $\lambda_i(L)$ are $q \times 1$ vector lag polynomials in nonnegative powers of L, and e_{it} is an idiosyncratic shock. That is, N series of data are assumed to be composed of two parts, common components, $\lambda_i(L)f_t$,and idiosyncratic errors e_{it}, $for\ each\ i$. Furthermore,

$$E(f_t e_{is}) = 0\ for\ all\ i, t, s, \tag{2}$$

and

$$E(e_{it} e_{js}) = 0\ for\ all\ i, j, t, s, i \neq j. \tag{3}$$

That is, the factors and idiosyncratic errors are assumed to be uncorrelated at all leads and lags and the idiosyncratic error terms are taken to be mutually uncorrelated at all leads and lags. Under this assumption, we call the DFM the exact DFM, which can be weakened by allowing some degree of serial correlation (the approximate DFM). Note that we do not impose parametric assumptions on idiosyncratic disturbances. In this nonparametric case, we can use the principal components method to estimate the factors and factor loadings after assuming identifying assumptions, as discussed in detail in the above papers.

Although maximum likelihood estimation is used with small datasets (see e.g., Stock and Watson 1989 and Quah and Sargent 1993), we are faced with an increasing number of parameters in large dataset environments. In such contexts, a simple way to proceed is to use principal component (see Stock and Watson 2006).

From equation (1), under the assumption that the lag polynomials has finite dimension, p, we can transform

the exact DFM into the static DFM as follows.

$$X_t = \Lambda F_t + e_t \ , \tag{4}$$

where $F_t = (f_t' f_{t-1}' ... f_{t-p+1}')'$ is $r \times 1$,where $r \leq pq$. Here r is the number of static factors. Λ is a factor loading matrix on the r static factors consisting of zeros and the coefficients of $\lambda_i(L)$. Since F_t consists of r static factors, we call equation (4) static DFM representation (Stock and Watson 2006). The static factors can be estimated as the principal components of the normalized data X_t.

Let us outline the estimation procedure. Following Stock and Watson (2006), let k ($k < \min\{N, T\}$) be an arbitrary number of factors, $N < T$, Λ be the $N \times k$ matrix of factor loadings, $(\Lambda_1, \Lambda_2, ..., \Lambda_N)'$, and F be a $T \times k$ matrix of factors $(F_1, F_2, ..., F_T)$. From equation (4), estimates of Λ and F_t are obtained by solving the following optimization problem :

$$\begin{aligned} V &= \min_{F,\Lambda} \frac{1}{T} \sum_{t=1}^{T} (X_t - \Lambda F_t)' (X_t - \Lambda F_t), \\ s.t \quad \Lambda'\Lambda &= I_k \end{aligned} \tag{5}$$

We treat $F_1, ..., F_T$ as fixed parameters to be estimated after normalizing Λ. Given $\widehat{\Lambda}$, the solution to equation (5) satisfy that $\widehat{F}_t = (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}' X_t$. Substituting this into equation (5) yields

$$\begin{aligned} V &= \min \frac{1}{T} \sum_{t=1}^{T} X_t'(I - \Lambda(\Lambda'\Lambda)^{-1}\Lambda')X_t \ \ s.t.\ \Lambda'\Lambda = I_k \\ &= \max tr((\Lambda'\Lambda)^{-\frac{1}{2}} \Lambda' \textstyle\sum_{XX} \Lambda (\Lambda'\Lambda)^{-\frac{1}{2}} \ \ s.t.\ \Lambda'\Lambda = I_k \\ &= \max \ \Lambda' \textstyle\sum_{XX} \Lambda \qquad\qquad\qquad\qquad s.t.\ \Lambda'\Lambda = I_k, \end{aligned}$$

where $\sum_{XX} = T^{-1} \sum_{t=1}^{T} X_t X_t'$. This optimization is solved by setting $\widehat{\Lambda}$ to the eigenvectors of matrix $X'X$ corresponding to its k largest eigenvalues. The estimator of factors is $\widehat{F}_t = \widehat{\Lambda}' X_t$.

For choosing the number of factors, we follow Bai and Ng (2002). After estimating $\widehat{\Lambda}$ and $\widehat{F}_t$, let $\hat{V}(k) = T^{-1}\sum_{t=1}^{T}(X_t - \widehat{\Lambda}\widehat{F}_t)'(X_t - \widehat{\Lambda}\widehat{F}_t)$ be the sum of squared residuals from regressions of X_t on the k factors and $IC(k) = \log(\hat{V}(k)) + k(\frac{N+T}{NT})\log(C_{NT}^2)$ be the information criterion where $C_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$. The consistent estimates of the true number of factors is then $\hat{k} = \arg\min_{0\leq k\leq \overline{k}} IC(k)$ where $\overline{k}$ is the maximum number of factors.

2.2 Mixed Frequency Factor Model

Unlike the above case wherein principal components is used to estimate the latent factors, now assume that the latent dynamics of a factor, m_t, from a mixed dataset, follows a zero-mean $AR(p)$ process. ADS (2009) and Aruoba and Diebold (2010) show that this seemingly simple latent factor captures the business cycle very well. The difference between the model outlined below and that specified in ADS (2009) is that we only use monthly data and quarterly data, unlike ADS (2009), where daily, weekly, and monthly series are exploited to construct m_t. Namely, we assume that the latent factor is updated every month. Let

$$m_t = \rho_1 m_{t-1} + \cdots + \rho_p m_{t-p} + e_t, \tag{6}$$

where e_t is white noise with unit variance. Here, we assume that there is a single factor in the economy. Thus, m_t is a scalar. This assumption can be generalized to include more factors or to allow for other models, including $ARMA$ and Markov switching models.

Let y_t^i denote the ith monthly economic or financial variable at month t, which depends *linearly* on m_t and possibly also on various exogenous variables $w_t^1, \ldots, w_t^k$ and/or lags of y_t^i, so the general measurement equation bridging y_t^i and latent factors is

$$\begin{aligned} y_t^i &= c_i + \beta_i m_t + \delta_{i1} w_t^1 + \cdots + \delta_{ik} w_t^k \\ &\quad + \gamma_{i1} y_{t-1}^i + \cdots + \gamma_{in} y_{t-n}^i + u_t^i, \end{aligned} \tag{7}$$

where the w_t^ks are exogenous variables and the u_t^i are contemporaneously and serially uncorrelated innovations. The variable y_t^i can be observed, or not. That is, if y_t^i is quarterly observed real GDP, say, then for the other two months y_t^i is not observed directly. To handle this problem systematically, following ADS (2009), we distinguish between stock and flow variables, observed data, and missing data.

Suppose that $\widetilde{y}_t^i$ denotes a stock variable observed at a lower (quarterly) frequency. At any time t, if y_t^i is observed, then $\widetilde{y}_t^i = y_t^i$. And if it is not observed, then $\widetilde{y}_t^i = NA$. Thus, the stock variable at time t is

$$\widetilde{y}_t^i = \begin{cases} y_t^i & , \; if \; y_t^i \; is \; observed \\ NA & , \; otherwise. \end{cases} \tag{8}$$

Combining equations (6) and equation (8), the measurement equation for a stock variable is

$$\widetilde{y}_t^i = \begin{cases} c_i + \beta_i m_{t-i} + \gamma_{i1}\widetilde{y}_{t-i}^i + \cdots + \gamma_{ip}\widetilde{y}_{t-n}^i + u_t^i & , \; if \; y_t^i \; is \; observed \\ NA & , \; otherwise. \end{cases} \tag{9}$$

Unlike a stock variable, a flow variable is observed at quarterly frequencies (e.g. real GDP), and can be interpreted as an intraperiod sum of the corresponding monthly values, so that a flow variable is defined as

$$\widetilde{y}_t^i = \begin{cases} \sum_{j=0}^{2} y_{t-j}^i & , \; if \; y_t^i \; is \; observed \\ NA & , \; otherwise. \end{cases} \tag{10}$$

Combining equations (6) and equation (10), the measurement equation for a flow variable is

$$\widetilde{y}_t^i = \begin{cases} c_i^* + \beta_i \sum_{i=0}^{2} m_t + \gamma_{i1}\widetilde{y}_{t-1}^i + \cdots + \gamma_{in}\widetilde{y}_{t-n}^i + u_t^{*i} & , \; if \; y_t^i \; is \; observed \\ NA & , \; otherwise. \end{cases} \tag{11}$$

Here, equation (6) is the state equation and equations (9) and (11) are the measurement equations. Together,

these equations constitute a state-space system. Given this fact, we can estimate mixed frequency factors via maximum likelihood using Kalman filtering and prediction error decomposition (see ADS (2009) for further details).

More specifically, the assumption on the factor dynamics and on the relation between the factor and the data can be represented by the following transition and measurement equations.

$$\begin{array}{ccccccc} \begin{bmatrix} m_{t+1} \\ m_t \\ m_{t-1} \end{bmatrix} & = & \begin{bmatrix} \rho & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} & \begin{bmatrix} m_t \\ m_{t-1} \\ m_{t-2} \end{bmatrix} & + & \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & [e_t] \\ \beta_{t+1} & = & T & \beta_t & + & R & e_t \end{array} \tag{12}$$

$$\begin{array}{ccccccccc} \begin{bmatrix} \widetilde{y}_t^1 \\ \widetilde{y}_t^2 \end{bmatrix} & = & \begin{bmatrix} \alpha_1 & 0 & 0 \\ \alpha_2 & \alpha_2 & \alpha_2 \end{bmatrix} & \begin{bmatrix} m_t \\ m_{t-1} \\ m_{t-2} \end{bmatrix} & + & \begin{bmatrix} c_1 & \gamma_1 & 0 \\ c_2 & 0 & \gamma_2 \end{bmatrix} & \begin{bmatrix} 1 \\ \widetilde{y}_{t-M}^1 \\ \widetilde{y}_{t-Q}^2 \end{bmatrix} & + & \begin{bmatrix} u_{1t} \\ u_{2t}^* \end{bmatrix} \\ y_t & = & Z & \beta_t & + & \Gamma & w_t & + & u_t \end{array}, \tag{13}$$

where

$$\begin{bmatrix} e_t \\ u_{1t} \\ u_{2t}^* \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_{u1}^2 & 0 \\ 0 & 0 & \sigma_{u2}^{*2} \end{bmatrix} \right),$$

and u_{2t}^* and σ_{2t}^{*2} signify, respectively, the measurement error and the variance thereof, in the case of flow variables. Also, in our simple setup, $\widetilde{y}_t^1$ represents a monthly stock variable and $\widetilde{y}_t^2$ represents a quarterly flow variable, so that $\widetilde{y}_t^2$ is empty (that is, NA in equations (9) or (11)) if the data was not released at time t). Let us now turn to the estimation of this system.

2.2.1 Kalman Filter and Signal Extraction

In order to illustrate the use of the Kalman filter, let us write the above two equations as follows:

$$\begin{aligned} Y_t &= Z\beta_t + \Gamma w_t + u_t \\ \beta_t &= T\beta_{t-1} + Re_t, \end{aligned}$$

where $u_t \sim N(0, Q)$ and $e_t \sim N(0, H)$. Y_t is a vector of observed variables, β_t is the latent state vector, which follows an AR(1) process, and w_t is a vector of exogenous variables. Under error normality, the Kalman filter can be used to estimate this system (see e.g., Anderson and Moore 1979, Harvey 1989, Kim and Nelson 1998). Following Kim and Nelson (1998), $Y_t \equiv [y_1, y_2, ..., y_t]$, $y_{t|t-1} = E[y_t|Y_{t-1}]$, $\eta_{t|t-1} = y_t - y_{t|t-1}$, $F_{t|t-1} = cov[\eta_{t|t-1}]$, $\beta_{t|t} = E(\beta_t|Y_t)$, $P_{t|t} = cov(\beta_t|Y_t)$, $\beta_{t|t-1} \equiv E(\beta_t|Y_{t-1})$, and $P_{t|t-1} = cov(\beta_t|Y_{t-1})$. If both variables are observed in month t, then we can use equation (12) and equation (13). Then, the Kalman filter consists of following six equations: For $t = 1, ...T$,

$$\beta_{t|t-1} = T\beta_t, \tag{14}$$

$$P_{t|t-1} = ZP_tZ' + RHR', \tag{15}$$

$$\eta_{t|t-1} = y_t - y_{t|t-1} = y_t - Z\beta_{t|t-1} - \Gamma w_t, \tag{16}$$

$$F_{t|t-1} = ZP_{t|t-1}Z' + Q, \tag{17}$$

$$\beta_{t|t} = \beta_{t|t-1} + P_{t|t-1}Z'F^{-1}_{t|t-1}\eta_{t|t-1}, \tag{18}$$

$$P_{t|t} = P_{t|t-1} - P_{t|t-1}Z'F^{-1}_{t|t-1}ZP_{t|t-1}. \tag{19}$$

If missing data exists in y_t, we only use monthly data as follows

$$y^*_t = Z^*\beta_t + \Gamma^*_t w_t + u^*_t, \tag{20}$$

$$u^*_t \sim N\ (0, H^*), \tag{21}$$

where $y_t^* = Wy_t,\ Z^* = WZ,\ \Gamma^* = W\Gamma,\ u_t^* = Wu_t$ and $H^* = WHW'$ for W =[1 0]. Note that y_t^* is a single datum. The vector W_t is defined to "choose" the equation relating observed data $\widetilde{y}_t^1$ from equation (13). In this case, the Kalman filter works the same as described above, after substituting y_t, Z, Γ and H for y_t^*, Z^*, Γ^* and H^*.

The above Gaussian state space model can be estimated using the Kalman filter. Moreover, maximum likelihood estimation can be carried out using the so-called prediction error decomposition method. Specifically, when two variables are observed at any time t, the (log) likelihood is incrementally increased by the following amount,

$$LogL = -\frac{1}{2}\sum_t [\log 2\pi + (\log |F_{t|t-1}| \ + \eta'_{t|t-1} F_{t|t-1}^{-1} \eta_{t|t-1})]. \tag{22}$$

When quarterly data is missing at time s, the likelihood is updated by

$$LogL = -\frac{1}{2}\sum_s [\log 2\pi + (\log |F_{t|t-1}^*| \ + \eta'_{t|t-1} F_{t|t-1}^{*-1} \eta_{t|t-1})], \tag{23}$$

where $F_{t|t-1}^* = Z^* P_{t|t-1} Z^{*\prime} + H^*,\ Z^* = WZ$ and $u_t^* = Wu_t$ for $W = [1\ 0]$.

We need to estimate the vector of factors, β_t, and the hyper-parameters, $\rho, \alpha_1, \alpha_2, c_1, c_2, \gamma_1, \gamma_2, \sigma_{u1}^2,\ \sigma_{u2}^{*2}$. Given t, the iteration from equation (14) to equation (19) in the Kalman filter is used to calculate the additional likelihood increment. Given initial conditions, the likelihood is built iteratively, from period $t = 1$ to T. Hyper-parameters are chosen to maximize the likelihood. For the initial choice, $\beta_{0|0}$ and $P_{0|0}$, assuming that factors are stationary, one can use the unconditional mean and covariance matrix of β_t. For complete details, see Kim and Nelson (1998). After estimation of the hyper-parameters, one simply plugs the estimates into the system and constructs estimates of the latent factor(s).

3 Empirical Methodology

3.1 Data

We construct real-time forecasts for the U.S. unemployment rate. We construct two kinds of factors, Diffusion Indexes (DI) and Mixed Frequency factors (MF). To construct diffusion indexes, we use the Stock and Watson (2005) dataset, which is extended through September 2009 by Kim and Swanson (2011). Specifically, the version used here has 143 monthly U.S. variables, from 1959:12 through 2009:9. To make all variables stationary, the series are transformed by taking logarithms and/or differencing, following the approach of Stock and Watson (2005). A description of the series and specific transformations used is given in the Appendix in Stock and Watson (2005). Constructed DIs and MFs are monthly frequency. In our forecasting experiments, we use the "first" and "second" DIs, defined by the magnitude of the eigenvalue associated with them.

We construct our mixed frequency factor using log differenced quarterly real GDP (from Federal Reserve Economics Data, Real Gross Domestic Product, 1 Decimal, ranging from first quarter of 1960 to second quarter of 2011) and log differenced monthly total nonfarm employment (from U.S. Bureau of Labor Statistics, National Current Employment Statistics ,ranging from January of 1960 to August of 2011). The mixed frequency factor is estimated using the dynamic factor model discussed above.

The release dates of real GDP and of total nonfarm employment are different. For example, the first release of real GDP in the first quarter 2011 was April 28, 2011, and it was revised several times. In case of employment data, BLS reports its preliminary estimates on the first Friday of the month. For the purpose of modelling parsimoniously, we assume that the release date of the data is the same as the first day of the month and of the quarter. However, note that if we specify a mixed frequency factor model based on daily data, we can rigorously match the release date and update information in real-time. In the literature, studies constructing real-time indexes, or nowcasting, in a timely manner include Giannone, Reichlin, and Small (2008), Aruoba, Diebold and Scotti (2009), Altissimo, Cristadoro, Forni, Lippi and Veronese (2010), Camacho and Pérez-Quirós

(2010), Mariano and Murasawa (2010) and Angelini, Camba-Mendéz, Giannone, Rünstler and Reichlin (2011).

3.2 Forecasting Methods

Before discussing our forecast models, we outline some details of our experiments. We divide the data set into two subsamples. The first subsample has T_1 observations, and the second subsample has T_2 observations, for a total of $T = T_1 + T_2$. Using first subsample, DIs and MF are estimated using the above models. In all cases, the number of AR lags is estimated using the SIC. Then, via OLS, we estimate a forecasting model that makes use of the estimated factors, and we construct an h-step ahead forecast. At $T_1 + 1$, we use $T_1 + 1$ observations to again construct the DIs and the MF. These are then in turn used to construct a forecasting model and to subsequently construct a new h step-ahead forecast. This procedure is continued, resulting in a sequence of T_2 ex-ante $h-$step ahead predictions. These predictions are then compared for various specifications, using RMSFEs (Root Mean Square Forecast Errors). In particular, predictions from a variety of models are compared with predictions from a benchmark AR(p) model (see Table 3).

The models considered are an AR(p) model, with lags selected using the SIC, and various factor augmented AR models in Table 3. We set $h = 1$. Note that the number of autoregressive lags can change according to the various different factors augmented models that are specified. Specifications of this type are suggested in Stock and Watson (2002a).

3.2.1 Diffusion Index Model

Following Stock and Watson (2002a), suppose that y_t is target variable to be forecasted. A DI forecasting equation is

$$\widehat{y}^h_{T+h|T} = \widehat{\alpha}_h + \sum_{j=1}^{m} \widehat{\beta}'_{hj} \widehat{F}_{T-j+1} + \sum_{j=1}^{p} \widehat{\gamma}_{hj} y_{T-j+1}, \tag{24}$$

where $\widehat{F}_t$ is a vector of k estimated DIs, m lags of the factors are included, h is the forecasting window, and p is the number of autoregressive lags. This is our generic model for forecasting y_{T+h} at time T using DIs and AR

terms, and is estimated using least squares. One can easily generalize this model to include a vector of exogenous variables. For the sake of parsimony, we report forecasts based on model estimated using only the first two DIs, as discussed above. In the Table 3, for example, models denoted by "DI" denote those estimated using first two DIs with no lags, that is, letting $m = 1, k = 2$, and $\widehat{\gamma}_{hj} = 0$. "1st DI" denotes forecasts using the first DI and "2nd DI" denotes forecasts using only the second DI. "DI-AR" includes both DIs as well as autoregressive terms with p lags chosen using the SIC, where $0 \leq p \leq 12$. As discussed in many papers, k can be also estimated using the criterion outlined in Bai and Ng (2002). Forecasts using this model are constructed in the following manner. At each recursive iteration, the panel dataset of stationary variables is standardized to have zero mean and unit variance. Then, the number of factors is fixed at either one or two and DIs are estimated using principal component method.

Suppose that m_t^i is a mixed frequency factor extracted using the mixed frequency dynamic factor model discussed above. (For the MF factors, we use two stationary data series, as discussed above -both series are log-differenced). In this case, we can simply generalize the above forecasting model as follows:

$$\widehat{y}_{T+h|T}^{h} = \widehat{\alpha}_h + \sum_{j=1}^{m} \widehat{\beta^f}'_{hj} \widehat{F}_{T-j+1} + \sum_{j=1}^{p} \widehat{\beta^m}'_{hj} \widehat{M}_{T-j+1} + \sum_{j=1}^{q} \widehat{\gamma}_{hj} y_{T-j+1}, \tag{25}$$

where $\widehat{M}_t$ is the vector of mixed frequency factors containing m_t^i s for all i, which is added to equation (24). We will call $\widehat{M}_t$ MF (Mixed Frequency) factors. The number of MF factors is predetermined. For example, the number of mixed frequency factors is two, if we assume that there are two factors and VAR dynamics of the latent factors. Of course, by using different datasets, we can extract different mixed frequency factors. For example, Aruoba and Diebold (2010) constructs their real activity index and inflation index using different sets of data. Forecasts using MF factors are constructed in the same way as in the case of pure DI and DI-AR models, except that an additional recursive step, wherein the MF is estimated, is included in the estimation procedure discussed above.

4 Empirical Results

Before presenting our results, consider the graphs of MF and DI factors presented in Figures 1 and 2. Figure 1 plots the MF factor and Figure 2 present two DIs. The MF factor is assumed to follow an AR(1) process and using quarterly real GDP (log-differences) and monthly nonfarm employment payroll (log-differences) from February 1960 to August 2011, as discussed above. And, DIs are constructed using 143 series from December 1959 to September 2009. The first and second DIs are presented along-side the unemployment rate in Figure 2, for the period February 1960 to September 2009. The first DI has qualitative properties that are very similar to the MF factor (compare Figure 1 with Figure 2). Namely, severe drops in the 1st DI and the MF factor coincide with the eight recession episodes over our sample period. This is particularly true for the first and second oil shock episodes and the 2008 crisis. However, the graph of the second DI exhibits quite different properties. During every recession, the second DI increases, which is consistent with unemployment movements during recessions.

To disentangle the components that make up our DI factors, we use the A(j) statistic in Bai and Ng (2006), which is applied in Armah and Swanson (2010) in order to construct "observable proxies" for diffusion indexes. This statistic can be used to "rank" variables in terms of their contribution to overall factor variation. We compare the estimated factors, DIs, and the 143 variables in the large data set. Tables 1 and 2 gather the results of this empirical exercise. Interestingly, the first DI depends crucially on real variables such as industrial production, as well as on nonfarm payroll, and capacity utilization; while the second DI is more closely tied to nominal bond yields and spreads.

The main results of our prediction experiment are summarized in Table 3, and although informative, should be taken only as an illustration of the methods discussed herein. The first column contains the abbreviation used to denote the prediction model. The models can be conveniently divided into two categories. First is our benchmark AR(p) model. The second is our set of factor augmented AR models. Numerical entries in the second column are relative RMSFEs (relative to the benchmark). Bold entries denote superior pointwise predictive performance, as compared with the benchmark. Evidently, our factor augmented models perform better than the benchmark, with the mixing model (i.e., the model that contains both a DI and a MF factor) performing best. Interestingly,

this model does not contain an AR component, suggesting that the factors are adequately capturing not only contemporaneous but also dynamic information useful for forecasting unemployment.

5 Concluding Remarks

We survey two varieties of latent factor model. The first is a convenient representation that allows for the use of simple principle components method for extracting estimates of latent factors from largescale datasets. The second type of model, estimated using the Kalman filter and smaller datasets, includes variables with differing observational frequencies. We find preliminary evidence that using a combination of factors constructed both of these ways as inputs into factor augmented forecasting equations yields improved predictions.

6 References

Altissimo, F., R. Cristadoro, M. Forni, M. Lippi and G.F. Veronese, (2010), New Eurocoin: Tracking Economic Growth in Real Time, *The Review of Economics and Statistics,* 92, 1024-1034.

Anderson, B.D.O. and J.B. Moore, (1979), *Optimal Filtering*, Prentice-Hall, Englewood Cliffs.

Angelini, E., G. Camba-Méndez, D. Giannone, L. Reichlin and G. Rünstler, (2011), Short-Term Forecasts of Euro Area GDP Growth, *The Econometrics Journal,* 14, 25-44.

Armah, N.A. and N.R. Swanson, (2010), Seeing inside the Black Box: Using Diffusion Index Methodology to Construct Factor Proxies in Large Scale Macroeconomic Time Series Environments, *Econometric Reviews*, 29, 476-510.

Armah, N.A. and N.R. Swanson, (2011), Some Variables are More Worthy than Others: New Diffusion Index Evidence on the Monitoring of Key Economic Indicators, *Applied Financial Economics*, 21, 43-60.

Aruoba, S.B., F.X. Diebold and C. Scotti, (2009), Real-Time Measurement of Business Conditions, *Journal of Business and Economic Statistics,* 27, 417-427.

Aruoba, S.B. and F.X. Diebold, (2010), Real-Time Macroeconomic Monitoring: Real Activity, Inflation, and Interactions, *The American Economic Review,* 100, 20-24.

Bai, J., (2003), Inferential Theory for Factor Models of Large Dimensions, *Econometrica,* 71, 135-171.

Bai, J. and S. Ng, (2002), Determining the Number of Factors in Approximate Factor Models, *Econometrica,* 70, 191-221.

Bai, J. and S. Ng, (2006), Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions, *Econometrica,* 74, 1133-1150.

Bai, J. and S. Ng, (2007), Determining the Number of Primitive Shocks in Factor Models, *Journal of Business and Economic Statistics,* 25, 52-60.

Bai, J. and S. Ng, (2008), Forecasting Economic Time Series Using Targeted Predictors, *Journal of Econometrics,* 146, 304-317.

Bai, J. and S. Ng, (2010), Instrumental Variable Estimation in a Data Rich Environment, *Econometric Theory,* 26, 1577-1606.

Bernanke, B.S. and J. Boivin, (2003), Monetary Policy in a Data-Rich Environment, *Journal of Monetary Economics,* 50, 525-546.

Bernanke, B.S., J. Boivin and P.S. Eliasz, (2005), Measuring the Effects of Monetary Policy: a Factor-Augmented Vector Autoregressive (FAVAR) Approach, *The Quarterly Journal of Economics,* 120, 387-422.

Boivin, J. and S. Ng, (2005), Understanding and Comparing Factor-Based Forecasts, *International Journal of Central Banking,* 1, 117-152.

Boivin, J. and S. Ng, (2006), Are More Data Always Better For Factor Analysis?, *Journal of Econometrics,* 132, 169-194.

Breitung, J. and J. Tenhofen, (2011), GLS Estimation of Dynamic Factor Models, *Journal of the American Statistical Association,* 106, 1150-1166.

Camacho, M. and G. Pérez-Quirós, (2010), Introducing the Euro-STING: Short-Term Indicator of Euro Area Growth, *Journal of Applied Econometrics*, 25, 663-694.

Camacho, M., G. Pérez-Quirós, and P. Poncela, (2012), Markov-Switching Dynamic Factor Models in Real Time, CEPR Discussion Papers, 8866.

Connor, G. and R. Korajczyk, (1986), Performance Measurement with the Arbitrage Pricing Theory, *Journal of Financial Economics,* 15, 373-394.

Connor, G. and R. Korajczyk, (1993), A Test for the Number of Factors in an Approximate Factor Model, *Journal of Finance,* 48, 1263-1291.

Connor, G. and R. Korajczyk, (1998), Risk and Return in an Equilibrium APT: Application of a New Test Methodology, *Journal of Financial Economics,* 21, 255-289.

Diebold, F.X. and R.S. Mariano, (1995), Comparing Predictive Accuracy, *Journal of Business and Economic Statistics,* 13, 253-263.

Doz, C., D. Giannone and L. Reichlin, (2011a), A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models, *The Review of Economics and Statistics,* forthcoming.

Doz, C., D. Giannone and L. Reichlin, (2011b), A Two-Step Estimator for Large Approximate Dynamic Factor Models Based on Kalman Filtering, *Journal of Econometrics,* 164, 188-205.

Forni, M., M. Hallin, M. Lippi and L. Reichlin, (2000), The Generalized Dynamic-Factor Model: Identification and Estimation, *The Review of Economics and Statistics* 82, 540-552.

Forni, M., M. Hallin, M. Lippi and L. Reichlin, (2005), The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting, *Journal of the American Statistical Association,* 100, 830-840.

Forni, M. and L. Gambetti, (2010), The Dynamic Effects of Monetary Policy: a Structural Factor Model Approach, *Journal of Monetary Economics,* 57, 203-216.

Forni, M., D. Giannone, M. Lippi and L. Reichlin, (2009), Opening the Black Box: Structural Factor Models with Large Cross Sections, *Econometric Theory,* 25, 1319-1347.

Forni, M. and M. Lippi, (2011), The General Dynamic Factor Model: One-Sided Representation Results, *Journal of Econometrics,* 163, 23-28.

Forni, M. and L. Reichlin, (1998), Let's Get Real: a Dynamic Factor Analytical Approach to Disaggregated Business Cycle, *The Review of Economic Studies,* 65, 453-474.

Giannone, D., L. Reichlin and D. Small, (2008), Nowcasting: the Real-Time Informational Content of Macroeconomic Data, *Journal of Monetary Economics,* 55, 665-676.

Ghysels, E., P. Santa-Clara and R.I. Valkanov, (2006), Predicting Volatility: Getting the Most Out of Returns Data Sampled at Different Frequencies, *Journal of Econometrics,* 131, 59-95.

Ghysels, E., A. Sinko and R.I. Valkanov, (2006), MIDAS Regressions: Further Results and New Directions, *Econometric Reviews,* 26, 53-90.

Jungbacker, B. and S.J. Koopman, (2008), Likelihood-Based Analysis of Dynamic Factor Models, Tinbergen Institute Discussion Paper.

Jungbacker, B., S.J. Koopman and M. van der Wel, (2011), Maximum Likelihood Estimation for Dynamic Factor Models with Missing Data, *Journal of Economic Dynamics and Control* 35, 1358-1368.

Hallin, M. and R. Liska, (2007), Determining the Number of Factors in the General Dynamic Factor Model, *Journal of the American Statistical Association,* 102, 603-617.

Harvey, A.C., (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Kim, C.-J. and C. Nelson, (1999), *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*, The MIT Press, Cambridge.

Kim, H.H. and N.R. Swanson, (2011), Forecasting Financial and Macroeconomic Variables Using Data Reduction Methods: New Empirical Evidence, Working Paper, Rutgers University.

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

Mariano, R.S. and Y. Murasawa, (2010), A Coincident Index, Common Factors, and Monthly Real GDP. *Oxford Bulletin of Economics and Statistics,* 72, 27-46.

Onatski, A., (2009), Testing Hypotheses about the Number of Factors in Large Factor Models, *Econometrica,* 77, 1447-1479.

Onatski, A., (2010), Determining the Number of Factors from Empirical Distribution of Eigenvalues, *The Review of Economics and Statistics,* 92, 1004-1016.

Quah, D. and T.J. Sargent, (1993), A Dynamic Index Model for Large Cross Sections, in: *Business Cycles, Indicators and Forecasting*, eds. James H. Stock, and Mark W. Watson, National Bureau of Economic Research.

Stock, J.H. and M.W. Watson, (1989), New Indexes of Coincident and Leading Economic Indicators, in: *NBER Macroeconomics Annual 1989, Volume 4*, edited by Olivier J Blanchard and Stanley Fischer, National Bureau of Economic Research.

Stock, J.H. and M.W. Watson, (1999). Forecasting Inflation, *Journal of Monetary Economics,* 44, 293-335.

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

Stock, J.H. and M.W. Watson, (2002b), Forecasting Using Principal Components from a Large Number of Predictors, *Journal of the American Statistical Association,* 97, 1167-1179.

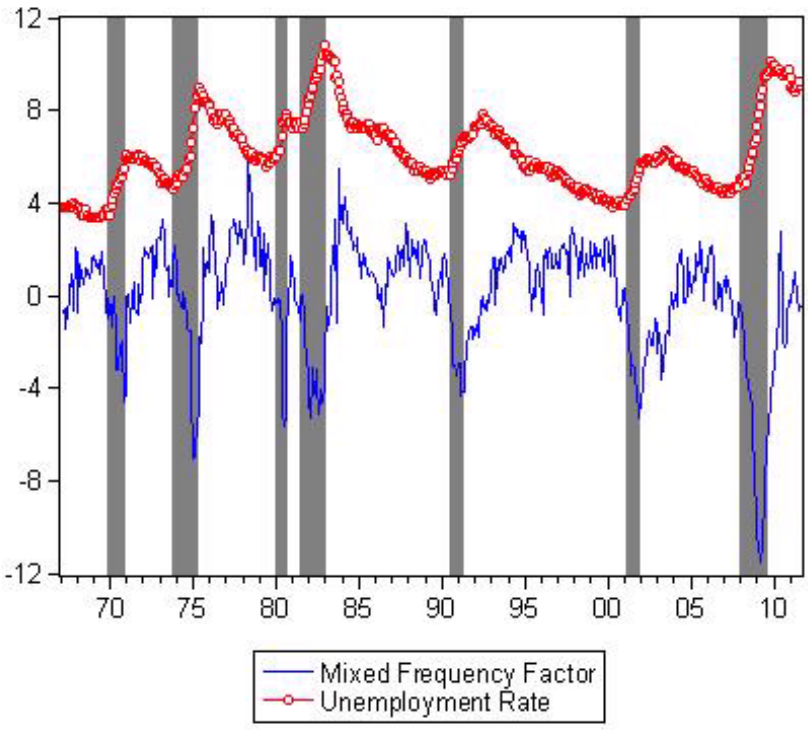
Stock, J.H. and M.W. Watson, (2005), Implications of Dynamic Factor Models for VAR analysis, *NBER Working Papers,* 11467.

Stock, J.H. and M.W. Watson, (2006), Macroeconomic Forecasting Using Many Predictors, in: *Handbook of Economic Forecasting,* edited by Clive W.J. Granger, Graham Elliott, and Allan Timmermann, Elsevier, Amsterdam.

Stock, J.H. and M.W. Watson, (2009), Forecasting in Dynamic Factor Models Subject to Structural Instability, in: *The Methodology and Practice of Econometrics: a Festschrift in Honour of Professor David F. Hendry*, edited by Jennifer L. Castle and Neil Shephard, Oxford University Press, Oxford.

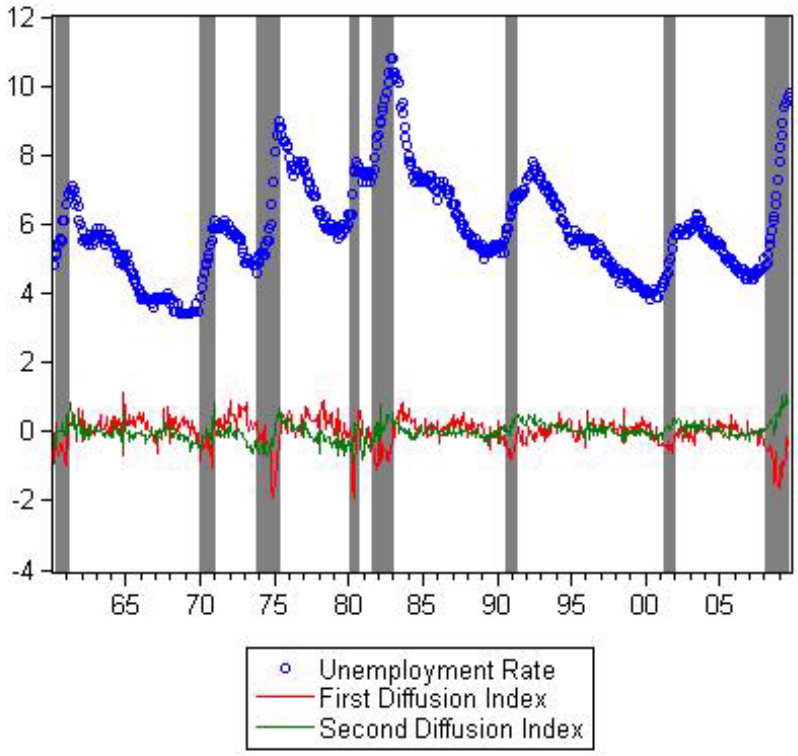
Stock, J.H. and M.W. Watson, (2011), Dynamic Factor Models, in: *Oxford Handbook of Forecasting,* edited by Michael P. Clements, and David F. Hendry, Oxford University Press, Oxford.

Figure 1: Mixed Frequency Factor and the Unemployment Rate



(*) Notes: The unemployment rate and a mixed frequency factor are plotted. Variables used to construct the MF factor include monthly (log differenced) total employment payroll and quarterly (log differenced) real GDP for the period 1960:2 - 2011:8. The MF factor has not been smoothed (see ADS (2009)). For further details, see Section 2.

Figure 2: First and Second Diffusion Indexes



(*) Notes: See notes to Figure 1.

Table 1: Top Observational Proxies for the 1st DI based on the A(j) Statistic

Ranking	1st Diffusion Index
1.	IP : Manufacturing
2.	IP : Total
3.	Nonfarm Payroll : Goods Pricing
4.	Nonfarm Payroll : Total Private
5.	Capacity Utilization : Manufacturing
6.	Employees on Nonfarm Payrolls : Total Nonfarm
7.	Employees on Nonfarm Payrolls : Manufacturing
8.	Employees on Nonfarm Payrolls : Durable Goods
9.	IP : Product
10.	IP : Materials

(*) Notes: This table ranks variables according to their contribution to the diffusion indexes used in the forecasting experiment reported in Sections 3 and 4. The A(j) statistic is from Bai and Ng (2005), and is examined in Armah and Swanson (2010).

Table 2: Top Observational Proxies for the 2nd DI based on the A(j) Statistic

Ranking	2nd Diffusion Index
1.	Spread : Moody's baa Corporate (% per annum) and Federal Funds Rate
2.	Spread : Moody's a Corporate (% per annum) and Federal Funds Rate
3.	Spread : Moody's Aaa Corporate (% per annum) and Federal Funds Rate
4.	Spread : Interest Rate on U.S. Treasury Constant Maturities, 10-year and Federal Funds Rate
5.	Spread : Interest Rate on U.S. Treasury Constant Maturities, 5-year and Federal Funds Rate
6.	Spread : Interest Rate on U.S. Treasury Bills, sec mkt, 3-month and Federal Funds Rate
7.	Spread : Interest Rate on U.S. Treasury Bills, sec mkt, 6-month and Federal Funds Rate
8.	Spread : Interest Rate on U.S. Treasury Constant Maturities, 1-year and Federal Funds Rate
9.	IP : Automotive
10.	Capital Utilization : Motor Vehicles and Parts

(*) Notes: See notes to Table 1.

Table 3: Out-of-Sample Forecasting Results*

Forecast Model	Relative Root MSFE (RMSFE)
AR	1
DI	**0.8724**
1st DI	**0.8693**
2nd DI	1.0787
DI-AR	1.0614
MF	**0.9199**
MF-AR	1.0681
MF-DI	**0.8710**
MF-DI-AR	1.1080
RMSFE, AR Model	0.1475

(*) Notes: Results of unemployment prediction experiments using various models both with and without latent factors are presented for a 1-month ahead forecast horizon. Models are listed in the left hand column. All numerical entries are root mean square forecast errors, relative to a benchmark AR(p), with lags selected using the Schwarz information criterion. Data used in model estimation and prediction construction are from the period 1960:2 - 2009:9. Model "DI" is a model that uses only 2 diffusion indexes as predictors. "1st DI" uses only the highest explanatory variance contributing diffusion index, while "2nd DI" uses only the the second highest diffusion index. Model "DI-AR" combines 2 diffusion indexes with an AR specification. Model "MF" is a forecasting model wherein only a single mixed frequncy factor is used as the explanatory variable. The rest of the models are permutations of those discussed above. For prediction experiment details, refer to Sections 3 and 4.