Forecasting Macroeconomic Variables using Collapsed Dynamic Factor Analysis

Brauning, F.U.; Koopman, S.J.

2012

document version
Early version, also known as pre-print

TI 2012-042/4
Tinbergen Institute Discussion Paper

Forecasting Macroeconomic Variables using Collapsed Dynamic Factor Analysis

Falk Brauning
Siem Jan Koopman

Faculty of Economics and Business Administration, VU University Amsterdam, and Tinbergen Institute.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Forecasting macroeconomic variables using collapsed dynamic factor analysis

Falk Bräuning and *Siem Jan Koopman* *
VU University Amsterdam, Department of Econometrics
Tinbergen Institute Amsterdam

April 18, 2012

Abstract

We explore a new approach to the forecasting of macroeconomic variables based on a dynamic factor state space analysis. Key economic variables are modeled jointly with principal components from a large time series panel of macroeconomic indicators using a multivariate unobserved components time series model. When the key economic variables are observed at a low frequency and the panel of macroeconomic variables is at a high frequency, we can use our approach for both nowcasting and forecasting purposes. Given a dynamic factor model as the data generation process, we provide Monte Carlo evidence for the finite-sample justification of our parsimonious and feasible approach. We also provide empirical evidence for a U.S. macroeconomic dataset. The unbalanced panel contain quarterly and monthly variables. The forecasting accuracy is measured against a set of benchmark models. We conclude that our dynamic factor state space analysis can lead to higher forecasting precisions when panel size and time series dimensions are moderate.

1 Introduction

Forecasting economic growth is a challenging task and it requires a good understanding of both economic theory and dynamic econometric modeling of macroeconomic and financial time series. The methodological development of economic forecasting is therefore still in process. In addition, the different crises since the collapse of Lehman Brothers in 2008 have given some ammunition to policy makers to review their methodology of forecasting

*Contact: f.brauning@vu.nl and s.j.koopman@vu.nl. We would like to thank James Mitchell, our discussant at the 8^{th} IIF Workshop 2011 on 'Forecasting the Business Cycle' at the Banque de France in Paris, for his valuable comments.

macroeconomic time series. This paper aims to provide a contribution to this debate by considering new methods for forecasting economic time series and by presenting empirical evidence for the U.S. economy.

We propose a collapsed dynamic factor analysis for the forecasting a target variable vector using information from many predictor variables. The dynamic factor model is collapsed by applying a dimension reduction on the high dimensional vector of predictors which we do not aim to forecast. A typical dimension reduction method that is used in this context is the principal components technique. We then analyse the target variable jointly with the collapsed vector of predictors by means of a multivariate unobserved components time series model that is represented as a linear Gaussian state space framework. The uncobserved components are all present in the equation for the target variable vector but a subset of the components are specifically linked to the collapsed vector of predictors, typically the principal components. Hence the information from the cross-section and time dimensions are accounted for simultaneously in the model. Due to the application of the dimension reduction technique, the model is far more parsimonious than the dynamic factor model specification for all series in the macroeconomic panel. It further allows for a flexible parametrization of the covariance structure in the idiosyncratic part of the target variable vector. The unknown parameters can be estimated using the method of maximum likelihood for which the loglikelihood function is evaluated by the Kalman filter. The proposed method can be implemented as a two-step procedure where principal component analysis produces first-step factor estimates that are in a second-step jointly modeled with the target variable in the state space framework. It combines principal component analysis and maximum likelihood estimation. The state space framework also allows for an unified and easy-to-implement treatment of time series analysis. Practically relevant issues such as the forecasting with mixed data frequency, nowcasting quarterly GDP from monthly macro panels, factor smoothing and the treatment of so-called jagged edges can be implemented straightforwardly.

Our modeling approach relates to several recent developments in dynamic factor analysis and in forecasting based on large panels of macroeconomic variables. The early contributions in the development of dynamic factor analysis have been recently reviewed by Stock and Watson (2006a), Breitung and Eickmeier (2006) and Bai and Ng (2008). Our approach is motivated by the diffusion indices of Stock and Watson (2002a, 2002b) . We adopt their use of principal components in the modeling of a vector of target variables. However, in our new modeling approach we analyse the target and the principal component variables simultaneously in a multivariate unobserved component time series model. A similar approach is taken by Doz, Giannone, and Reichlin (2011) who propose a two-step estimation method that is based on a dynamic factor model with the factor loadings set equal to the eigenvectors associated with a set of principal components. In the first step, the principal components are computed and its dynamic properties are estimated by means of a vector autoregressive

model. In the second step, factor estimates and forecasts are obtained from Kalman filter methods applied to a model with the eigenvectors as factor loadings and with autoregressive coefficient matrices for the factors set equal to those estimated from the principal components. Doz, Giannone, and Reichlin (2011) provide the asymptotic properties of the Kalman filter estimates and apply the model to nowcasting quarterly GDP with monthly variables that are released in a non-synchronized dating scheme. The Kalman filter estimates have exploited the factor dynamics and it is therefore expected that the resulting factor estimates are more efficient compared to principal components estimates.

Our approach is distinctive from the approach of Doz, Giannone, and Reichlin (2011) since we adopt a simultaneous model for the target variable, the principal components and the unobserved dynamic factors, and we estimate all parameters in this parsimonious model by the method of maximum likelihood. In this setting we aim to capture all cross-sectional and time information in an optimal way. The idiosyncratic parts of the target vector series are specified explicitly and estimated jointly with the common factors. Hence we prevent the problem that factors estimated from a large macroeconomic panel might be irrelevant to the forecasting target.

A Monte Carlo experiment illustrates the forecasting performance of the collapsed model and compares it with forecasts from models which include principal components or other factor estimates as predictors. We find that the collapsed factor model outperforms standard methods in terms of mean square forecast errors, specifically for models where irrelevant factors for the target series are included and where macroeconomic panels have only small time and cross-sectional dimensions. These are cases which seem in particularly relevant for small countries where macro panels are less extensive and for institutions where the means to maintain large databases for forecasting are not available.

The remainder of the paper is organized as follows. In the next section we review principal components analysis and dynamic factor state space analysis. We also establish notation and discuss some methods in detail for future references. In section 3, we introduce our new method of a collapsed dynamic factor analysis. Issues related to forecasting with mixed data frequencies, nowcasting quarterly variables using panels of monthly macroeconomic time series, factor smoothing and treating data with jagged edges are discussed in detail. The results of a Monte Carlo study is presented in section 4. Empirical evidence is given in section 5. We find that our feasible methods do not compromise in its forecasting performance when compared with other methods. In most cases the collapsed dynamic factor model outperforms benchmark and competitor models. Section 6 concludes.

2 Review of dynamic factor models

2.1 Static and dynamic model specifications

Assume that observation x_{it} is for variable i observed at time t with $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The variables are part of a large panel with cross-section dimension N and time series length T. Observations in the panel may be missing. Dynamic factor models assume that the time series properties for each variable in the panel rely on common and idiosyncratic unobserved components. The common component is typically represented by a small set of dynamic factors which drive the bulk of correlations both in the cross-section and time series dimensions. The static representation of a factor model is given by

$$x_t = \Lambda F_t + e_t, \qquad t = 1, \ldots, T, \tag{1}$$

where $x_t = (x_{1t}, \ldots, x_{Nt})'$, F_t is a vector of $r << N$ common factors, $\Lambda = (\lambda_1', \ldots, \lambda_N')'$ is an $N \times r$ matrix with λ_i the $1 \times r$ vector of factor loadings for variable i, and $e_t = (e_{1t}, \ldots, e_{Nt})'$ is a vector of idiosyncratic components or errors. In the remainder of the paper we assume that prior to the analysis all time series in the panel are demeaned and transformed to stationarity. We also do not consider the inclusion of regressors and constant terms in the model although the inclusion of these terms is a straightforward extension of our analysis. Since the processes for F_t and e_t are uncorrelated at all leads and lags, the covariance matrix of x_t is given by

$$\Sigma_x = \Lambda \Sigma_F \Lambda' + \Sigma_e.$$

The factors and factor loadings are only identified up to a pre-multiplication of an invertible matrix. We assume the usual identifying restriction $\Sigma_F = I_r$ so that $\Sigma_x = \Lambda\Lambda' + \Sigma_e$. This restriction fixes the scaling of factors up to their multiplication by an orthonormal matrix.

In the dynamic factor model literature, different names are used to refer to the model (1) when different structures are imposed on the variance matrix Σ_e. When Σ_e is diagonal, the idiosyncratic components are cross-sectionally uncorrelated and the covariances between the variables in the panel are due the common factors. This model is usually referred to as an exact factor model. When a limited degree of covariance structure in e_t is allowed, the model is often referred to as an approximate factor model; see, for example, Chamberlain and Rothschild (1983). The relation between x_t and F_t in (1) is static in the sense that current x_t is related to current F_t. The dynamic properties for x_t are introduced by imposing time series processes for F_t and e_t. A popular dynamic specification for F_t is given by the vector autoregressive (VAR) process for some lag order p_F as

$$F_t = \Phi_1 F_{t-1} + \ldots + \Phi_{p_F} F_{t-p_F} + u_t, \qquad t = 1, \ldots, T, \tag{2}$$

where Φ_i is the $r \times r$ autoregressive coefficient matrix for $i = 1, \ldots, p_F$ and u_t is the r-dimensional disturbance vector for which each element is identically and independently distributed (i.i.d.) with zero mean and variance matrix Σ_u. The idiosyncratic vector component is also assumed to follow a VAR processes of order p_e and represented by

$$e_t = \Gamma_1 e_{t-1} + \ldots + \Gamma_{p_e} e_{t-p_e} + v_t, \tag{3}$$

where Γ_i is an $N \times N$ diagonal autoregressive coefficient matrix for $i = 1, \ldots, p_e$ and v_t is an N-dimensional i.i.d. random vector with mean zero and variance matrix Σ_v. Both dynamic specifications can be modified in many different ways. For example, moving average terms can be added to the specifications.

In contrast with the static representation, the dynamic representation of a factor model is given by

$$x_t = \Lambda_0 F_t + \Lambda_1 F_{t-1} + \ldots + \Lambda_s F_{t-s} + e_t, \tag{4}$$

where Λ_j is an $N \times r$ matrix of factor loadings for $j = 0, 1, \ldots, s$ for some maximum lag order s and disturbance vector e_t has the same properties as in (1). This model relates x_t to both current and lagged factors $F_t, F_{t-1}, \ldots, F_{t-s}$. By adding moving arverage terms to this specification so that the lag order s is effectively infinite, the model is referred to as a generalized dynamic factor model; see, for example, Forni, Hallin, Lippi, and Reichlin (2000, 2005) and Forni and Lippi (2001).

The dynamic representation of the factor model can also be represented by the static specification (1). We define

$$F_t^+ = \begin{pmatrix} F_t \\ F_{t-1} \\ \vdots \\ F_{t-s} \end{pmatrix}, \qquad \Lambda^+ = [\Lambda_0, \Lambda_1, \ldots, \Lambda_s],$$

so that we can represent (4) as $x_t = \Lambda^+ F_t^+ + e_t$ which is the static model (1). Hence we can adopt the static form in many general settings and for many purposes including macroeconomic forecasting. The estimation of the parameters and of the factors in the static model are straightforwardly carried out using time domain methods. We therefore will concentrate on the static factor representation in the developments below.

2.2 Principal component analysis

An important motivation for a dynamic factor analysis based on the static model (1) is to extract and to forecast the common factors F_t given the data $x_1, \ldots, x_T$. Stock and Watson (2002a, 2002b) propose a non-parametric method based on principal component analysis

in the time domain; see also the references in Breitung and Eickmeier (2006). Denote the matrix of unobserved factors $F = (F_1, \ldots, F_T)$, factor loadings matrix $\Lambda = (\lambda_1', \ldots, \lambda_N')'$ and $N \times T$ data matrix $X = (x_1, \ldots, x_T)$ where $x_t = (x_{1t}, \ldots, x_{Nt})'$ for $t = 1, \ldots, T$. The static model (1) can then be expressed in matrix form as

$$X = \Lambda F + E, \qquad E = (e_1, \ldots, e_T).$$

The principal component approach is based on the minimization of the nonlinear objective function

$$(NT)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N} (x_{it} - \lambda_i' F_t)^2 = (NT)^{-1} \mathrm{tr}[(X - \Lambda F)'(X - \Lambda F)] = (NT)^{-1} \mathrm{tr}(E'E),$$

with respect to F and Λ. The minimizing values are denoted by $\hat{F}$ for F and $\hat{\Lambda}$ for Λ. It follows that

$$\hat{F} = (\Lambda'\Lambda)^{-1}\Lambda' X.$$

We concentrate out $\hat{F}$ from the objective function. The problem reduces to the maximization of $\mathrm{tr}(XX'\Lambda\Lambda')$ with respect to Λ and subject to the identification restriction $\Lambda'\Lambda/N = I_r$. Hence the estimates for Λ are obtained by applying a principal component analysis to XX'; we obtain

$$\hat{\Lambda} = U = (U_1, \ldots, U_r), \tag{5}$$

where U_j is the eigenvector corresponding to one of the r largest ordered eigenvalues u_j of XX' for $j = 1, \ldots, r$. The resulting principal component estimate for F is then given by

$$\hat{F}_{PC} = \hat{\Lambda}' X = U' X. \tag{6}$$

Stock and Watson (2002b) refer to these factor estimates as diffusion indices. An alternative analysis can be based on a principal component analysis applied to $X'X$ which is computationally more convenient when $N > T$.

The principal components are consistent estimates of the true factor F when both T and N go to infinity; see Stock and Watson (2002a). However, the principal component estimates do not take account of possible heteroskedasticity and cross-correlation in the idiosyncratic component of the model. In other words, the principal component factor estimates do not account for the data properties as specified by the parametric model (2) and are therefore in general inefficient. The principal component estimates are only efficient when we consider an exact factor model with homoscedastic idiosyncratic components.

When the ratio of the variation due to common factors versus the variation due to the idiosyncratic factors (signal-to-noise ratio) is small in the static model (1), Onatski (2009)

shows that the principal component estimates of the factors and loadings are inconsistent. This is a relevant problem for a large set of variables in a macroeconomic panel of time series. Onatski (2009) derive a formula for the extent of inconsistency; it can be used to carry out an asymptotic correction for the estimates. Karoui (2008) argue that if the cross-section dimension N and the time dimension T are both large and both of the same magnitude, the eigenvectors provide inaccurate estimates for Λ. We may conclude that the principal component estimates are in many cases inefficient and even erroneous in the way the estimates of F are extracted from the data. However, it is also true that the method of principal components is simple to implement.

2.3 Dynamic factor state space analysis

The static factor model specification (1) – (3) can be cast in state space form and the Kalman filter with its related methods can be used for its analysis. For example, the Kalman filter and smoother enable the extraction of the latent factors from the time series panel when the parametric model is known. Unknown parameters can be estimated by the method of maximum likelihood in which the loglikelihood function is evaluated by the Kalman filter and is maximised numerically using an appropriate quasi-Newton method. The state space framework provides a unified approach to time series analysis for almost all linear Gaussian models; see, for example, Harvey (1989) and Durbin and Koopman (2001) for textbook treatments. The general linear state space model is based on observation and state equations as given by

$$x_t = Z_t\alpha_t + \varepsilon_t, \qquad \alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \tag{7}$$

respectively, with observation vector x_t, latent state vector α_t, and i.i.d. disturbance vectors $\varepsilon_t \sim \mathrm{N}(0, \Sigma_\varepsilon)$ and $\eta_t \sim \mathrm{N}(0, \Sigma_\eta)$. The system matrices Z_t, T_t and R_t, together with the disturbance covariance matrices Σ_ε and Σ_η, are deterministic and they completely determine the dynamic statistical properties of x_t.

The dynamic factor model can be represented in state space form. As an example we take the model with a vector autoregressive process of order p_F, VAR(p_F), for the common factor component and an i.i.d. sequence for the idiosyncratic component. We set the state vector as $\alpha_t = \left(F_t', F_{t-1}', \ldots, F_{t-p_F}'\right)'$. The state space representation of model (1) and (2) is as follows. The observation equation is given by

$$x_t = \begin{pmatrix} \Lambda & 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} F_t \\ F_{t-1} \\ \vdots \\ F_{t-p_F} \end{pmatrix} + \varepsilon_t, \tag{8}$$

with $\Sigma_\varepsilon = \Sigma_e$, the covariance matrix of the idiosyncratic term. The state transition equation

in (7) is given by the companion form of the VAR(p_F) process, we have

$$\begin{pmatrix} F_{t+1} \\ F_t \\ F_{t-1} \\ \vdots \\ F_{t-p_F+1} \end{pmatrix} = \begin{pmatrix} \Phi_1 & \Phi_2 & \dots & \Phi_{p_F-1} & \Phi_{p_F} \\ I_r & 0 & \dots & 0 & 0 \\ 0 & I_r & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & I_r & 0 \end{pmatrix} \begin{pmatrix} F_t \\ F_{t-1} \\ \vdots \\ F_{t-p_F+1} \\ F_{t-p_F} \end{pmatrix} + \begin{pmatrix} I_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} u_t, \qquad (9)$$

with $\Sigma_\eta = \Sigma_u$. Similar state space representations can be provided for other dynamic factor models. The dynamic specification of the factor model in (4) is obtained in the state space form (8) and (9) by replacing in the Z_t matrix of (8), the matrix Λ with Λ_0 and the zero block matrices with Λ_j for $j = 1, \dots, s$. A similar representation of the dynamic factor model is given by Forni, Hallin, Lippi, and Reichlin (2000).

The Kalman filter and smoother evaluates the minimum mean square linear estimator (MMSLE) of the facors F_t, together with its variance, conditional on all observations, that is

$$\mathbb{E}(\alpha_t|x_1, \dots, x_T), \qquad \mathbb{V}\text{ar}(\alpha_t|x_1, \dots, x_T), \qquad t = 1, \dots, T, \qquad (10)$$

where F_t is a part of α_t. The resulting MMSLE for all factors $F = (F_1', \dots, F_T')'$ is denoted by $\hat{F}_{KFS}$. The Kalman filter can also be used for the computation of the h-step ahead MMSLE forecast of F_t, that is

$$\mathbb{E}(\alpha_{T+h}|x_1, \dots, x_T), \qquad \mathbb{V}\text{ar}(\alpha_{T+h}|x_1, \dots, x_T), \qquad h = 1, 2, \dots, \qquad (11)$$

where F_t is a part of α_t. The MMSLE property is specific to the estimates from the Kalman filter and smoother and is not shared by the principal components estimates. For a correctly specified model, the maximum likelihood parameter estimates are consistent and efficient under weak regularity conditions. The analysis also provides specification tests to verify the correct specification of the model. In particular, diagnostic tests for normality and serial correlation are widely used.

Parameter estimation by maximum likelihood, specifically for those in the loading matrix, becomes a heavy task when the panel dimension increases as the number of parameters growth in N. State space methods have therefore typically been used for models with small and moderate panel dimensions N; see, for example, Engle and Watson (1981) and Mariano and Murasawa (2003). Hence most applications focus therefore on principal component estimation; see the references in the survey of Bai and Ng (2008).

Recent approaches in dynamic factor analysis, however, has moved towards maximum likelihood estimation via Kalman filtering and smoothing; see Jungbacker and Koopman (2008), Kapetanios and Marcellino (2009), Doz, Giannone, and Reichlin (2011). Jungbacker and Koopman (2008) propose to transform the observation equation into a lower dimension

which leads to a computationally efficient approach to parameter and factor estimation. Kapetanios and Marcellino (2009) suggest the use of the so-called sub-space algorithm for parameter estimation with the purpose of avoiding high-dimensional computations. In a recent development, Doz, Giannone, and Reichlin (2011) estimate the factors, the loadings and the VAR coefficients of a dynamic factor model in two steps. In the first step, parameter estimation is based on a principal component analysis. In the second step, the factors are re-estimated by the Kalman filter and smoother. They show that their two-step procedure leads to consistent factor estimates.

3 Collapsed dynamic factor analysis

3.1 Incorporating a target variable in a factor model

Our aim is to reduce the dimension of the observation vector in a way that preserves the information on common components and the idiosyncratic dynamics of a specific subset of all time series in the panel. The dimension reduction should enable maximum likelihood estimation by means of the Kalman filter in a feasible fashion even when the panel dimension is large. The main motivation for our approach is forecasting of macroeconomic time series when many variables are present. In applied econometrics, we are often only interested in forecasting a selection of target series while the future values of other series are of minor or no interest beyond the information they contain to forecast the target series. Examples of target series are gross domestic product (GDP) and inflation. Such variables provide a summary of the state of the economy.

Suppose that y_t denotes the L-dimensional target variable. Its dynamic properties are represented by a set of time series components that we collect in the state vector α_{yt}. The initial time series model for y_t under consideration can be given by its state space form

$$y_t = \Lambda_{yy}\alpha_{yt} + \varepsilon_{yt}, \qquad \alpha_{y,t+1} = T_{yy}\alpha_{yt} + R_{yy}\eta_t, \tag{12}$$

where the same assumptions apply as for the general linear state space model (7). The system matrices Λ_{yy}, T_{yy} and R_{yy} are fixed and their elements rely partially on unknown parameters. The initial model for y_t can be regarded as a time series model in which no other information is used for the forecasting of y_t than its own past realisations. With the aim to improve the forecast precision of the model for y_t, we augment the observation vector y_t in its state space representation (12) to $(y_t', x_t')'$ where the $N \times 1$ vector x_t represents a large panel of macroeconomic variables, with $L << N$. The time series vector x_t is assumed to be generated by the static factor model (1) and where the factor F_t is modelled by the general linear state space model (7) where F_t is a part of the state vector α_t. We propose to

analyse the augmented observation vector $(y_t', x_t')'$ using the model

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{bmatrix} \Lambda_{yy} & \Lambda_{yx} \\ 0 & \Lambda_{xx} \end{bmatrix} \begin{pmatrix} \alpha_{yt} \\ F_t \end{pmatrix} + \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{xt} \end{pmatrix}, \qquad \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{xt} \end{pmatrix} \sim \mathrm{N}\left(0, \begin{bmatrix} \Sigma_{yy} & 0 \\ 0 & \Sigma_{xx} \end{bmatrix}\right), \quad (13)$$

for $t = 1, \ldots, T$, where Λ_{xx} represents Λ and ε_{xt} represents e_t in (1). The additional factor coefficient matrix in the augmented model is Λ_{yx} which is assumed to be fixed and unknown. The factor loading matrix Λ_{yx} allows the information in x_t to be used for the modeling and forecasting of the target variable y_t through the factor F_t. The variable x_t loads uniquely on F_t while variable y_t loads both on the state α_{yt} and F_t. We restrict the model to have no interaction between x_t and α_{yt} since we are focused on the generation of the optimal forecasts for y_t only.

It follows almost immediately that the augmented model can be represented by the linear state space model (7). Specifically, we can express the augmented model as

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{bmatrix} \Lambda_{yy} & \Lambda_{yx} & 0 & \ldots \\ 0 & \Lambda_{xx} & 0 & \ldots \end{bmatrix} \begin{pmatrix} \alpha_{yt} \\ \alpha_{xt} \end{pmatrix} + \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{xt} \end{pmatrix}, \qquad \alpha_{xt} = \begin{pmatrix} F_t \\ F_{t-1} \\ \vdots \end{pmatrix}, \quad (14)$$

where we can formulate the dynamic process for the state vector $\alpha_t = (\alpha_{yt}', \alpha_{xt}')'$ by the state updating equation $\alpha_{t+1} = T_t\alpha_t + R_t\eta_t$ in (7) or, more specifically,

$$\begin{pmatrix} \alpha_{y,t+1} \\ \alpha_{x,t+1} \end{pmatrix} = \begin{bmatrix} T_{yy} & 0 \\ 0 & T_{xx} \end{bmatrix} \begin{pmatrix} \alpha_{yt} \\ \alpha_{xt} \end{pmatrix} + \begin{bmatrix} R_{yy} \\ R_{xx} \end{bmatrix} \eta_t, \qquad \eta_t \sim \mathrm{N}(0, \Sigma_\eta), \quad (15)$$

for $t = 1, \ldots, T$, with reference to equation (12) and the use of obvious notation.

An illustration of our general modeling framework can be presented for a single target variable y_t, that is $L = 1$. We assume that an appropriate dynamic model for y_t is given by the trend-cycle decomposition model of Harvey and Jaeger (1993) as given by

$$y_t = \mu_t + \psi_t + \varepsilon_{yt}, \qquad \varepsilon_{yt} \sim \mathrm{N}(0, \sigma^2_{\varepsilon,y}), \quad (16)$$

where μ_t is the trend component, ψ_t is the cycle component and ε_{yt} is the i.i.d. disturbance with mean zero and variance $\sigma^2_{\varepsilon,y}$, for $t = 1, \ldots, T$. Different dynamic specifications for the nonstationary trend component and for the stationary stochastic cycle component can be considered and are discussed in detail by Harvey (1989). All specifications can be cast into the state space form (12) with the state vector α_{yt} containing the variables μ_t and ψ_t and matrix Λ_{yy} selecting these variables into the observation equation for y_t. The forecasting performance of this univariate time series model may be improved by including a set of dynamic factors that can be constructed from a large macroeconomic time series panel.

Hence, for this purpose we can consider the joint model

$$y_t = \mu_t + \psi_t + \Lambda_{yx} F_t + \varepsilon_{yt}, \quad x_t = \Lambda_{xx} F_t + \varepsilon_{xt}, \quad \varepsilon_{yt} \sim \mathrm{N}(0, \sigma^2_{\varepsilon,y}), \quad \varepsilon_{xt} \sim \mathrm{N}(0, \Sigma_{\varepsilon,x}), \quad (17)$$

which is directly in the form of the augmented observation equation (13). From an analysis based on this model and for a real data set, we can assess whether the inclusion of the factor F_t in the model for y_t improves the forecasting performance of the model. As N increases, the number of parameters increases rapidly and estimation becomes a heavy computational task. We therefore aim to develop a feasible estimation procedure that remains feasible even when the panel size N increases to high numbers.

3.2 Collapsing the dynamic factor model

To avoid a dynamic factor analysis based on the high-dimensional $N \times 1$ vector x_t and to avoid the estimation of all unknown parameters including those in Λ_{xx}, we introduce the collapsed dynamic factor model. We carry out a transformation based on the same eigenvector space that is used for the computation of principal components. In particular, we pre-multiply the observation equation (13) by the transformation matrix

$$P = \begin{bmatrix} I_L & 0 \\ 0 & A \end{bmatrix} \quad (18)$$

where matrix A is an $r \times N$ matrix where r is the dimension of the number of factors in F_t. Since $r < N$, the transformed vector $A(y_t', x_t')'$ has a reduced dimension. We obtain

$$\begin{pmatrix} y_t \\ Ax_t \end{pmatrix} = \begin{bmatrix} I_L & \Lambda_{yx} \\ 0 & A\Lambda_{xx} \end{bmatrix} \begin{pmatrix} \alpha_{yt} \\ F_t \end{pmatrix} + \begin{pmatrix} \varepsilon_{yt} \\ A\varepsilon_{xt} \end{pmatrix}, \quad (19)$$

for $t = 1, \ldots, T$. The transition equation for the dynamic specifications of both α_{yt} and F_t remains unchanged.

The choice of matrix A in (18) determines the information loss we are willing to accept from the dimension reduction of x_t. We typically want to find a transformation that preserves as much information as is relevant for the forecasting of y_t. Here we consider the suggestion of Stock and Watson (2002a, 2002b) by taking $A = U'$ where U is defined in (5) as the $N \times r$ matrix of eigenvectors used for the construction of the principal components for XX'. The principal component estimates are given by $\hat{F}_{PC,t} = U'x_t$ for $t = 1, \ldots, T$. We assume loosely that $F_t \approx \hat{F}_{PC,t}$ which we can formalize by imposing that $F_t = \hat{F}_{PC,t} +$ error and

$$U'\Lambda_{xx} = I_r.$$

When N is sufficiently large, we may expect that the approximation becomes more accurate. By defining the transformation matrix as

$$P_{PC} = \begin{bmatrix} I_L & 0 \\ 0 & U' \end{bmatrix}$$

we obtain the collapsed dynamic factor model as

$$\begin{pmatrix} y_t \\ \hat{F}_{PC,t} \end{pmatrix} = \begin{bmatrix} I_L & \Lambda_{yx} \\ 0 & I_r \end{bmatrix} \begin{pmatrix} \alpha_{yt} \\ F_t \end{pmatrix} + \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{PC,t} \end{pmatrix}. \tag{20}$$

where $\varepsilon_{PC,t} = U' (\varepsilon_{xt} - [U - \Lambda_{xx}]F_t)$ for $t = 1, \ldots, T$. The disturbance $\varepsilon_{PC,t}$ includes the error due to replacing U for the loading matrix Λ_{xx}. We expect this error to be small since $U'\Lambda_{xx} \approx I_r$. Hence we suppose that $\varepsilon_{PC,t}$ is approximately i.i.d. with zero mean and a variance matrix that can be partly derived from its construction. However, we will treat $\mathbb{V}\text{ar}(\varepsilon_{PC,t})$ as an unknown variance matrix.

Since it is encountered in many studies based on both economic theory and statistical data analysis that the number of r common factors is relatively small, the collapsed dynamic factor model relies on small dimensions, even when x_t represents a large time series panel of economic and financial variables. Due to the parsimonious structure with many zeros and ones in the system matrices, only a few parameters need to be estimated by the method of maximum likelihood. The use of more parsimonious models are preferred in empirical studies; it is often concluded from forecasting comparison studies that parsimonious models outperform large econometric models with many parameters in their forecasting accuracy.

The collapsed dynamic factor model representation in (20) emphasizes that the principal components in $\hat{F}_{PC,t}$ are explicitly treated as an errors-in-variables problem. The errors-in-variables are made explicit in the model and their magnitude is determined by the variance matrix of $\varepsilon_{PC,t}$. Clearly, when the variance of some linear function of $\varepsilon_{PC,t}$ reduces to zero, a part of the error is non-existent and F_t in the state vector α_{xt} is partly observed. More specifically, when all variances of $\varepsilon_{PC,t}$ are zero, F_t is equal to $\hat{F}_{PC,t}$ and the model for y_t is closely related to the dynamic model used in the second step of a principal component analysis. The variance matrix of $\varepsilon_{PC,t}$ is estimated by the method of maximum likelihood in our analysis. Finally, it is straightforward to represent the collapsed dynamic factor model in state space form. For completeness, we present the model here as

$$\begin{pmatrix} y_t \\ \hat{F}_{PC,t} \end{pmatrix} = \begin{bmatrix} I_L & \Lambda_{yx} & 0 & \ldots \\ 0 & I_r & 0 & \ldots \end{bmatrix} \begin{pmatrix} \alpha_{yt} \\ \alpha_{xt} \end{pmatrix} + \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{PC,t} \end{pmatrix}, \tag{21}$$

where the state vector α_{xt} is defined in (14) and the updating equations for the state vectors α_{xt} and α_{yt} are given in (15).

The collapsed dynamic factor approach offers a compromise between the principal component analysis of section 2.2 and the parametric dynamic factor state space analysis of section 2.3. We keep the state space framework but we reduce the observation and parameter vector dimensions substantially with a minimum loss of information. The Kalman filter and smoother method is used for the signal extraction of the factors F_t using the target series y_t and the principal components $\hat{F}_{PC,t}$ for $t = 1, \ldots, T$. Hence the factor estimates are established jointly by the macroeconomic time series panel x_t (via the principal components $\hat{F}_{PC,t}$) and by the target variable y_t. A particular feature of our model is that the principal components in $\hat{F}_{PC,t}$ are treated as dependent variables that are exclusively associated with the factors F_t. At the same time, the factors F_t are added to the time series model equation for the target variable y_t. In the approach of Stock and Watson (2002a, 2002b), the principal components are placed directly in the model equation for the target variable. Since principal components may be regarded as (possibly) noisy estimates of the factors in F_t, we expect a more effective use of the principal components when we model them explicitly as noisy indicators of the factors, and jointly with the target variable y_t.

We can summarize our procedure as a two-step procedure. We first carry out a principal component analysis for dimension reduction of the large panel of macroeconomic time series. This step circumvents the maximum likelihood method for estimating many parameters in a large dimensional state space model. Hence we avoid an infeasible or expensive exercise which also may erode the forecasting performance. In the second step we model the principal components jointly with the target variable y_t in a small-scale dynamic factor state space model a small number of parameters. The unknown parameters are estimated by the method of maximum likelihood in a standard manner. Our reliance on the state space framework in the second step allows us to keep the benefits of the use of Kalman filter methods for signal extraction, forecasting, diagnostic checking of residuals, and handling of missing observations.

The collapsed factor model is related to the two step procedure for factor estimation by Doz, Giannone, and Reichlin (2011). In their approach the target variable y_t is supposed to be part of x_t and the focus is then on x_t that is analysed by the static factor model (1) and (2). The first step also carries out a principal component analysis but with the purpose of estimating the parameters of the model. A VAR model is fitted on the principal component estimates $\hat{F}_{PC,t}$ to provide estimates for the VAR coefficient matrices and the variance matrix of the VAR disturbance vector in (2). The loading matrix Λ in (1) is set equal to U which is consistent with a principal component analysis; compare equation (5). The estimate of the disturbance variance matrix Σ_ε is set to $I_N - UU'$ where we assume that x_t is standardized. In the second step, the state space model (8) and (9) is considered with all unknown matrices replaced by their estimated counterparts from the first step. The Kalman filter and smoother method is used to obtain the in-sample estimates and the out-of-sample forecasts of F_t. This

two step procedure is used in empirical studies; see Giannone, Reichlin, and Small (2008) and Banbura and Rünstler (2011). Conceptually the two procedures have similarities since they are both grounded in a principal component analysis. However, a distinguishing feature is the role of the target variable y_t which is diminished in the procedure of Doz, Giannone, and Reichlin (2011). In our procedure we have the flexibility to design a specific dynamic model for y_t that may already provide good forecasts for y_t. All unknown parameters in our model are estimated by the method of maximum likelihood in the second step.

3.3 Forecasting with the collapsed dynamic factor model

The forecasting of macroeconomic time series using information from many predictors can be carried in the context of a collapsed dynamic factor model. Many practical issues such as forecasting with mixed frequency data, missing observations, nowcasting and backcasting can be easily accommodated in our proposed framework. For expositional purposes we focus on the case where we need to forecast a single target variable y_t and where we typically consider the model (17) with trend μ_t fixed at some constant μ, cycle ψ_t specified as the autoregressive process $\psi_t = \phi\psi_t + \kappa_t$, with i.i.d. disturbance $\kappa_t \sim \mathrm{N}(0, \sigma_\kappa^2)$, and with F_t specified as the VAR(p_F) process (2), that is $F_t \sim \mathrm{VAR}(p_F)$ with $p_F = 1$. The model for y_t is then given by $y_t = \mu + \psi_t + \Lambda_{yx} F_t + \varepsilon_t$, with $\varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$, for $t = 1, \ldots, T$. Models for a vector of target variables with more unobserved components and other dynamic specifications are regarded as straightforward generalizations.

For given values of the parameters in this model, the minimum mean square linear error h-step ahead forecast $\hat{y}_{T+h}$ at time T is given by

$$\begin{aligned}
\hat{y}_{T+h} &= \mathbb{E}(y_{T+h}|\Omega_T) \\
&= \mu + \mathbb{E}(\psi_{T+h}|\Omega_T) + \Lambda_{yx}\,\mathbb{E}(F_{T+h}|\Omega_T) \\
&= \mu + \phi^h\,\mathbb{E}(\psi_T|\Omega_T) + \Lambda_{yx}\Phi_1^h\,\mathbb{E}(F_T|\Omega_T),
\end{aligned} \tag{22}$$

where $\Omega_t = \{y_t, y_{t-1}, \ldots, y_1, x_t, x_{t-1}, \ldots, x_1\}$ is the observed information set at time t and where $\mathbb{E}(a_T|\Omega_T)$ is obtained from the Kalman filter for $a = \psi, F$. The last two equalities are only valid for the specific model under consideration. Different model specifications and different parameter choices, possible due to the use of different estimation methods, will lead to different forecasts $\hat{y}_{T+h}$. This approach to forecasting applies to any model that is represented in state space form. Hence it applies to the dynamic factor state space model (14) and to the collapsed dynamic factor model (21) where in both cases the state updating equations are given by (15). When considering the collapsed dynamic factor model, the information set Ω_t is effectively reduced to $\Omega_t = \{y_t, y_{t-1}, \ldots, y_1, \hat{F}_{PC,t}, \hat{F}_{PC,t-1}, \ldots, \hat{F}_{PC,1}\}$.

The forecasting practice in a principal component analysis is mainly developed by Stock and Watson (2002b). When the principal component estimates $\hat{F}_{PC,t}, \ldots, \hat{F}_{PC,T}$ are com-

puted, the h-step ahead forecast for the target variable is based on the autoregressive model with explanatory variables as given by

$$y_{t+h} = \mu^h + \varphi_0^h y_t + \ldots + \varphi_{p_y}^h y_{t-p_y} + \beta_0^h \hat{F}_{PC,t} + \ldots + \beta_{k_y}^h \hat{F}_{PC,t-k_y} + u_{yt}, \tag{23}$$

for some integers p_y and k_y, with regression coefficients $\mu^h, \varphi_0^h, \ldots, \varphi_{p_y}^h, \beta_0^h, \ldots \beta_{k_y}^h$, and i.i.d. disturbances u_{yt}, for $t = 1, \ldots, T-h$. The regression coefficients can be estimated by least squares. The regression coefficients have superscripts h to emphasize that for each forecasting horizon h, a new least squares regression takes place with new regression coefficient estimates. The actual forecast $\hat{y}_{T+h}$ is then computed as

$$\hat{y}_{T+h}^{SW} = \hat{\mu}^h + \hat{\varphi}_0^h y_T + \ldots + \hat{\varphi}_{p_y}^h y_{T-p_y} + \hat{\beta}_0^h \hat{F}_{PC,T} + \ldots + \hat{\beta}_{k_y}^h \hat{F}_{PC,T-k_y}, \tag{24}$$

for some positive integer h. This forecasting approach based on principal components is widely used; see, for example, Marcellino, Stock, and Watson (2003), Breitung and Eickmeier (2006) and the references therein. Many different forecasts for y_t can be constructed in this way: we only require to replace the principal component estimate $\hat{F}_{PC,t}$ by another factor estimate in (24). For example, we can use the factor estimates from the procedure of Doz, Giannone, and Reichlin (2011) as discussed in section 3.2.

The methods of forecasting in a principal component analysis and in a dynamic factor state space analysis are clearly different but in both cases the forecasts $\hat{y}_{T+h}$ are weighted linear functions of the available data $y_1, \ldots, y_T$ and $x_1, \ldots, x_T$; see the discussion in Durbin and Koopman (2001, Chapter 4). The different constructions of the forecasting observation weights are based on different optimum criteria. In the sections below we will investigate in more detail which approaches are providing the best weights for forecasting when many predictors are available. These questions will be analysed in both a Monte Carlo simulation study and in an empirical study based on different data sets.

3.4 Forecasting density

The MMSLE forecasting equations (22) and (24) only provide point forecasts. For many purposes of practical interest, the point forecasts are not sufficiently informative. On the other hand, it may not be feasible or desirable to generate forecasts from other statistical procedures which are based on loss functions motivated from economic theory and principles. As an alternative, economic researchers often focuses on density forecasting to produce prediction intervals of the h-step ahead variable y_{T+h}. Density forecasting may provide a more comprehensive picture of future economic developments.

Under the assumption of normally distributed disturbances in the state space model, the variance (or the standard deviation) for the MMSLE point forecasts provides, together with

the mean (or the point forecast), a complete and well-defined forecast density. For example, when we assume that the parameter estimates are equal to their true parameter values, we obtain the conditional variance of y_{T+h} as

$$\begin{aligned}\mathbb{V}\mathrm{ar}(y_{T+h}|\Omega_T) &= \mathbb{V}\mathrm{ar}(\mu + \psi_{T+h} + \Lambda_{yx}F_{T+h} + \varepsilon_{T+h}|\Omega_T) \\ &= \mathbb{V}\mathrm{ar}(\psi_{T+h}|\Omega_T) + \Lambda_{yx}\,\mathbb{V}\mathrm{ar}(F_{T+h}|\Omega_T)\Lambda_{yx}' + \mathbb{V}\mathrm{ar}(\varepsilon_{T+h}) \\ &= \mathbb{V}\mathrm{ar}(\psi_{T+h}|\Omega_T) + \Lambda_{yx}\,\mathbb{V}\mathrm{ar}(F_{T+h}|\Omega_T)\Lambda_{yx}' + \sigma_\varepsilon^2, \end{aligned} \tag{25}$$

where $\mathbb{V}\mathrm{ar}(a_{T+h}|\Omega_T)$ is obtained from the Kalman filter for $a = \psi, F$. We notice that for a linear Gaussian model, we have $\mathbb{V}\mathrm{ar}(\hat{y}_{T+h}) = \mathbb{V}\mathrm{ar}(y_{T+h}|\Omega_T)$. Hence the forecasting density is given by

$$y_{T+h}|\Omega_T \sim \mathrm{N}\left(\hat{y}_{T+h|T},\, \mathbb{V}\mathrm{ar}[y_{T+h}|\Omega_T]\right), \qquad h = 1, 2 \ldots.$$

The prediction interval for $\hat{y}_{T+h}$ can be obtained in the usual way as for any normal random variable. In practical applications, parameters are estimated from the observations and we need to accommodate the parameter estimation errors when the forecasting density is constructed. For general state space models, to account for parameter uncertainty, bootstrap methods can be used for the adjustments of $\hat{y}_{T+h|T}$ and $\mathbb{V}\mathrm{ar}(y_{T+h}|\Omega_T)$; see Stoffer and Wall (1991) for a more detailed discussion. In our approach, forecasting in a dynamic factor analysis, whether it is based on a full or a collapsed dynamic factor model, is carried out via a state space model and therefore the forecasting density can be obtained straightforwardly using the Kalman filter.

When applying the forecasting procedures of Stock and Watson (2002a, 2002b) and Doz, Giannone, and Reichlin (2011), but also for other step-wise forecast procedures, density forecasting requires additional adjustments. In the first step of such procedures, the factors are typically estimated by principal components. In the second step, these estimates are used as regressors or as indicators in a model with the aim to analyse of to forecast the target series. The forecasting intervals or densities are then constructed from this model. Although the principal components are consistent estimates of the factors, Bai and Ng (2006) show that the effect of estimated regressors should be taken into account for the construction of forecast intervals, unless T/N goes to infinity. They also provide a procedure to correct for the use of principal components as regressors. Although the collapsed dynamic factor model also relies on principal component estimates, the Bai and Ng (2006) correction procedure is not needed here. The principal components are treated as dependent variables in our approach. They have been the result of a data transformation. The consequences of this transformation for the model are incorporated in its specification. Hence the correction procedure does not apply to a collapsed dynamic factor analysis.

3.5 Mixed frequency observations

Most forecasting applications with many predictors are based on time series panels with variables that are sampled at different frequencies. An example of practical relevance is the forecasting of quarter-on-quarter growth (usually measured in percentages) of gross domestic product (GDP) which is sampled at a quarterly frequency. Many of the predictors in the model are month-on-month changes of variables which are sampled at monthly frequencies. For expositional purposes, we explore this example further and notice that generalisations of this case to other, mostly higher, frequencies are straightforward extensions of the solution presented below.

When we consider GDP as a stock variable, we can simply treat the quarterly series as a monthly series where the observations in the first two months of each quarter are treated as missing and are ignored for the updating equations in the Kalman filter. However, GDP is a flow variable and we introduce autocorrelation by construction when we model a quarterly flow variable at a monthly frequency. We define the unobserved (annualized) month-on-month GDP growth as

$$y_t = 12 \times \log(\text{GDP}_t \ / \ \text{GDP}_{t-1}), \qquad t = 1, \ldots, T.$$

We consider a dynamic factor model with GDP as the target series. We further define

$$y^Q_{3k} = 4 \times \log(\text{GDP}_{3k} \ / \ \text{GDP}_{3(k-1)}),$$

as the observed (annualized) quarter-on-quarter GDP growth for quarter $k = 1, \ldots, \lfloor T/3 \rfloor$ where $\lfloor b \rfloor$ is the largest integer value that is smaller than b. We can express the observed quarterly GDP growth as the average of the monthly growth rates, that is

$$y^Q_{3k} = \frac{1}{3}(y_{3k-1} + y_{3k-2} + y_{3k-3}), \qquad k = 1, \ldots, \lfloor T/3 \rfloor. \tag{26}$$

We observe the monthly variable $y^Q_t \equiv y^Q_{3k}$ in the third month of each quarter and we treat y^Q_t as a missing value for the other months.

The dynamic factor model can incorporate the relation between the quarterly obvserved variable and the underlying monthly flow variable that is not observed. In a monthly model, the dynamic processes are formulated in the monthly frequency. The implied dynamics for a quarterly variable defined as (26) is then accounted for by the model explicitly. We therefore keep the latent monthly GDP variable in the model and assume that it is modeled as in our earlier example : $y_t = \mu + \psi_t + \Lambda_{yx} F_t + \varepsilon_t$ with constant μ, stationary AR(1) process ψ_t, and i.i.d. disturbance ε_t. We take $\mu = 0$ for expositional purposes. Furthermore we introduce a cumulator variable to obtain the quarterly variable from the monthly rates; see Harvey (1989, Chapter 8) for a more detailed discussion. The aggregation will be established by the

latent cumulator variable y_t^C that is generated by

$$y_{t+1}^C = \delta_t y_t^C + \frac{1}{3} y_{t+1}, \qquad \delta_t = \begin{cases} 0, & t = 3k, \\ 1, & \text{otherwise}, \end{cases} \qquad y_1^C = \frac{1}{3} y_1, \qquad k = 1, \ldots, \lfloor T/3 \rfloor,$$

for $t = 1, \ldots, T$, where t refers to a month and k to a quarter. We assume that y_t^Q is only observed at the third month of each quarter.

The dynamic factor state space representation is constructed as follows. The observation equation becomes

$$\begin{pmatrix} y_t^Q \\ \hat{F}_t^{PC} \end{pmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & I_r \end{bmatrix} \begin{pmatrix} y_t \\ y_t^C \\ \psi_t \\ F_t \end{pmatrix} + \begin{pmatrix} 0 \\ \varepsilon_{PC,t} \end{pmatrix}, \tag{27}$$

and the transition equation for the extended state vector is given by

$$\begin{bmatrix} -1/3 & 1 & 0 & 0 \\ 1 & 0 & -1 & -\Lambda_{yx} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & I_r \end{bmatrix} \begin{pmatrix} y_{t+1} \\ y_{t+1}^C \\ \psi_{t+1} \\ F_{t+1} \end{pmatrix} = \begin{bmatrix} 0 & \delta_t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \phi & 0 \\ 0 & 0 & 0 & \Phi_1 \end{bmatrix} \begin{pmatrix} y_t \\ y_t^C \\ \psi_t \\ F_t \end{pmatrix} + \begin{pmatrix} 0 \\ \varepsilon_{t+1} \\ \kappa_t \\ u_t \end{pmatrix}, \tag{28}$$

for $t = 1, \ldots, T$. These observation and state equations can be placed in the usual state space form (7) with observation x_t, state α_t and disturbance η_t given by

$$x_t = \begin{pmatrix} y_t^Q \\ \hat{F}_t^{PC} \end{pmatrix}, \qquad \alpha_t = \begin{pmatrix} y_t \\ y_t^C \\ \psi_t \\ F_t \end{pmatrix}, \qquad \eta_t = \begin{pmatrix} \varepsilon_{t+1} \\ \kappa_t \\ u_t \end{pmatrix},$$

and with system matrices

$$Z_t = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & I_r \end{bmatrix}, \quad T_t = \begin{bmatrix} 0 & 0 & \phi & \Lambda_{yx}\Phi_1 \\ 1 & \delta_t & \phi/3 & \Lambda_{yx}\Phi_1/3 \\ 0 & 0 & \phi & 0 \\ 0 & 0 & 0 & \Phi_1 \end{bmatrix}, \quad R_t = \begin{bmatrix} 1 & 1 & \Lambda_{yx} \\ 1/3 & 1/3 & \Lambda_{yx}/3 \\ 0 & 1 & 0 \\ 0 & 0 & I_r \end{bmatrix},$$

for $t = 1, \ldots, T$. We notice that the transition matrix T_t is time-varying due to the inclusion of the deterministic indicator variable δ_t. Further discussions of mixed frequencies with empirical illustrations in the context of dynamic factor models are provided by Mariano and Murasawa (2003), Mitchell, Smith, Weale, Wright, and Salazar (2005) Proietti and Moauro

(2006) and Banbura and Rünstler (2011).

An alternative approach to the analysis of time series with mixed frequencies is the mixed data sampling regression (MIDAS) method as proposed by Ghysels, Santa-Clara, and Valkanov (2006). The MIDAS method provides linear projections without specifying the dynamics of the regressors. When the model is correctly specified and the parameters are known, the Kalman filter is superior to MIDAS by construction. Otherwise, it is under investigation whether MIDAS or the state space method is superior; see the study of Bai, Ghysels, and Wright (2011) where both MIDAS and state space methods are considered. They show under which conditions the methods are identical and provide evidence that the Kalman filter is slightly more accurate.

3.6 Missing values, unbalanced panels and rigged edges

In a time series panel with many macroeconomic variables, it is likely that different series start and end at different times. Such a panel is often referred to as an unbalanced panel. Also specific sections in the time series may not be available. Hence many missing values can be present in a macroeconomic panel. The treatment of unbalanced panels and missing values is straightforward in the the state space framework because missing observation can be handled explicitly in the Kalman filter; see Durbin and Koopman (2001, Chapter 4). In a collapsed dynamic factor analysis we need to distinguish between missings in the target variable y_t which is treated directly in the model, and missings in the panel vector x_t which we transform into a set of principal components. We need to account for the missing values when computing the principal components. A specific expectation-maximization algorithm can be used as described in Stock and Watson (2002b). Alternatively, the Kalman filter and smoother can be adopted before the collapse takes place. We can consider the dynamic factor model for x_t and take $\Lambda_{xx} = U$. Estimation of the other parameters, including those for the vector autoregressive process for F_t, can be based on the principal components from a balanced subsample of X; this procedure is suggested by Doz, Giannone, and Reichlin (2011). The principal component estimates can be extracted from this "estimated" balanced panel.

The two methods also provide solutions for the treatment of data sets with jagged or rigged edges which are due to unsynchronized data releases and publication lags. We can consider jagged edges as a specific structure of missing values in the time series panel. The incorporation of publication lags is specifically relevant in the "nowcasting" of GDP. The aim of nowcasting is to use information such as market expectations that can be generated from survey and financial data when forecasting quarterly GDP in the current quarter. Survey and financial data is often available weeks before other indicators from, for example, the national accounts are released; see the discussions in Giannone, Reichlin, and Small (2008) and Banbura and Rünstler (2011).

4 Monte Carlo study and finite sample properties

In this section we investigate how the out-of-sample forecast precision from a collapsed dynamic factor model compares to the forecasts from other factor-based approaches including those of Stock and Watson (2002b) and Doz, Giannone, and Reichlin (2011). We carry out an extensive Monte Carlo study to address the forecasting performances for different cross-section (N) and time series (T) dimensions, and for different forecasting horizons h. We study in more detail the collapsed dynamic factor model. For example, we want to assess whether the collapsed dynamic factor model can mitigate weak factors (relative to the idiosyncratic part) and erode forecasting performances.

In the Monte Carlo study, we compare the forecast performances with those of random walk (RW) and univariate autoregressive (AR) models which are typical benchmarks in many forecasting studies. Furthermore, we consider the forecast results from the principal component analysis of Stock and Watson (2002b, SW), the factor estimates in a Kalman filter and smoother analysis of Doz, Giannone, and Reichlin (2011, DGR), and our collapsed dynamic factor analysis (CFM). We consider data generating processes (DGP) that also have been used in Stock and Watson (2002b) and Doz, Giannone, and Reichlin (2011). The DGP allows for a certain degree of cross- and serial-correlation in the idiosyncratic terms as well as the factor dynamics. The design of this DGP reflects features of empirical relevance and is given by

- $x_{it} = \sum_{j=1}^{r} \lambda_{ij} f_{jt} + \xi_{it}$, for $i = 1, \ldots, N$ and $t = 1, \ldots, T$; in vectors $X_t = \Lambda F_t + \xi_t$;
- $\lambda_{ij} \sim \mathrm{N}(0, 1)$, for $i = 1, \ldots, N$ and $j = 1, \ldots, r$;
- $A(L)F_t = u_t$ with $u_t \sim N(0, (1-\rho^2)I_r)$ and for $i, j = 1, \ldots, r$; we take the (i, j) element of $A(L)$ as $a_{ij}(L) = 1 - \rho L$, if $i = j$ and 0 otherwise;
- $D(L)\xi_t = v_t$ with $v_t \sim N(0, Q)$ and for $i, j = 1, \ldots, N$; we take the (i, j) element of $D(L)$ as $d_{ij}(L) = 1 - dL$, if $i = j$ and 0 otherwise; the elements of variance matrix Q are given by

$$Q_{ij} = \sqrt{q_i q_j}\tau^{|i-j|}(1 - d^2), \qquad i, j = 1, \ldots, N,$$

  with

$$q_i = \frac{c_i}{1 - c_i} \sum_{j=1}^{r} \lambda_{ij}^2, \qquad c_i \sim U([C, 1 - C]).$$

A special case of this DGP is obtained when we set $\tau = 0$ so that no cross-correlations appear in the idiosyncratic terms corresponding to an exact factor model. If in addition $d = 0$ and $C = 0.5$, we have a so-called spherical, exact static factor model for which the principal component estimates are efficient. The coefficient c_i determines the noise-to-signal ratio, this is the ratio between the variance of the idiosyncratic component and the variance

of x_{it}. In our simulation study this ratio is uniformly distributed with an average of 0.5 unless specified differently.

Table 1: Monte Carlo study for exact factor model

This table presents the MSEs of 5000 h-step ahead forecasts for $h = 1, 2, 3, 6, 12$ and different sample sizes N and T. Parameter values of DGP model reflect exact factor model ($\rho = 0.9$, $d = 0$, $\tau = 0$, $C = 0.1$). Different forecast models are random walk (RW), autoregression of order 2 (AR), the principal component SW method, the smoothed factor DGR method, and our collapsed dynamic factor model (CFM). The smallest MSE for each experiment is highlighted.

	T = 50					T = 100				
	h = 1	h = 2	h = 3	h = 6	h = 12	h = 1	h = 2	h = 3	h = 6	h = 12
	N = 10					N = 10				
RW	1.2432	1.3577	1.4864	1.7301	2.0553	1.1847	1.2755	1.4087	1.6410	1.9047
AR	0.8384	0.8989	0.9429	1.0018	1.1681	0.8124	0.8780	0.9369	0.9980	1.1044
SW	0.7791	0.8761	0.9510	1.0885	1.3387	0.7278	0.7972	0.8846	0.9879	1.1826
DGR	0.7818	0.8793	0.9548	1.0788	1.3126	0.7385	0.8091	0.8959	0.9910	1.1712
CFM	0.7782	0.8645	0.9288	1.0287	1.2101	0.7298	0.7961	0.8884	0.9716	1.0954
	N = 50					N = 50				
RW	1.2930	1.4020	1.5154	1.7783	2.1457	1.1656	1.2975	1.4281	1.6858	1.9327
AR	0.8369	0.9086	0.9288	1.0217	1.1567	0.7978	0.8284	0.9101	1.0021	1.1038
SW	0.7343	0.8489	0.9226	1.1190	1.3398	0.6731	0.7377	0.8519	1.0235	1.1869
DGR	0.7346	0.8497	0.9223	1.1124	1.3210	0.6749	0.7388	0.8530	1.0236	1.1808
CFM	0.7340	0.8187	0.8674	1.0258	1.1913	0.6685	0.7319	0.8364	0.9605	1.0885
	N = 100					N = 100				
RW	1.2729	1.4282	1.4971	1.7676	2.1076	1.1455	1.2736	1.3875	1.5737	1.8809
AR	0.8608	0.9015	0.9366	1.0155	1.1920	0.8008	0.8740	0.9260	0.9983	1.1006
SW	0.7478	0.8502	0.9406	1.1214	1.3661	0.6719	0.7635	0.8487	1.0060	1.1880
DGR	0.7483	0.8498	0.9392	1.1167	1.3568	0.6724	0.7649	0.8501	1.0059	1.1833
CFM	0.7426	0.8294	0.8868	1.0385	1.2225	0.6720	0.7542	0.8312	0.9570	1.0925

Dynamic factor models are typically used for short-term forecasting and therefore we present the results for forecast horizons $h = 1, 2, 3, 6, 12$. We generate data for a given set of parameters with two common factors, compute the mean square errors (MSE), and repeat this exercise 5,000 times for each experiment. We abstract from model misspecification by taking the correct number of lags and factors for all models and all experiments.

In Table 1 we present the results for an exact factor model with homoskedastic idiosyncratic errors as the DGP. The smallest MSEs (among the 5 different forecasts of RW, AR, SW, DGR and CFM) for different values of T, N and h are highlighted. We confirm the results reported by Doz, Giannone, and Reichlin (2011) who find that the MSEs from SW and DGR are similar for all reported T/N combinations. However, we find that the MSEs from the collapsed dynamic factor model (CFM) in most cases are smaller compared to all benchmarks although differences are small in relative terms. To investigate whether, for small and moderate T/N ratios, principal components are noisy factor estimates, we compute the average and median of the estimated noise-to-signal ratio $\mathbb{V}\text{ar}(\varepsilon_{PC,t})$ / $\mathbb{V}\text{ar}(F_t)$ over all simulations. In the case of $T = 10$ and $N = 50$, the median of the noise-to-signal ratio estimates is close to zero so that the factor estimates from CFM are close to the principal

Table 2: Monte Carlo study for approximate factor model

This table presents the MSEs of 5000 h-step ahead forecasts for $h = 1, 2, 3, 6, 12$ and different sample sizes N and T. Parameter values of DGP model reflect approximate factor model ($\rho = 0.9$, $d = 0.5$, $\tau = 0.5$, $C = 0.1$). Different forecast models are random walk (RW), autoregression of order 2 (AR), the principal component SW method, the smoothed factor DGR method, and our collapsed dynamic factor model (CFM). The smallest MSE for each experiment is highlighted.

	T = 50					T = 100				
	h = 1	h = 2	h = 3	h = 6	h = 12	h = 1	h = 2	h = 3	h = 6	h = 12
	N = 10					N = 10				
RW	1.2437	1.3549	1.4865	1.7319	2.0549	1.2129	1.3074	1.3731	1.6827	1.9132
AR	0.8391	0.8978	0.9442	1.0038	1.1667	0.8278	0.8824	0.9006	1.0509	1.0965
SW	0.7854	0.8801	0.9520	1.0849	1.3169	0.7431	0.8151	0.8637	1.0613	1.1587
DGR	0.7882	0.8799	0.9541	1.0839	1.3261	0.7394	0.8125	0.8604	1.0620	1.1630
CFM	0.7842	0.8480	0.9122	1.0090	1.1809	0.7427	0.8131	0.8494	1.0252	1.0795
	N = 50					N = 50				
RW	1.2886	1.4294	1.5189	1.8234	2.1605	1.1898	1.2774	1.3776	1.6066	1.9485
AR	0.8563	0.9444	0.9534	1.0650	1.2152	0.8062	0.8566	0.9113	1.0133	1.1309
SW	0.7471	0.8790	0.9214	1.1457	1.4089	0.6676	0.7609	0.8450	1.0101	1.2020
DGR	0.7457	0.8779	0.9216	1.1516	1.4197	0.6659	0.7601	0.8400	1.0077	1.2081
CFM	0.7438	0.8598	0.8797	1.0666	1.2483	0.6687	0.7512	0.8224	0.9665	1.1180
	N = 100					N = 100				
RW	1.2594	1.3542	1.4745	1.7283	2.0128	1.2199	1.3241	1.4505	1.6683	2.0644
AR	0.8472	0.8631	0.9357	1.0278	1.1260	0.8228	0.8461	0.9200	0.9799	1.1339
SW	0.7257	0.8228	0.9306	1.1120	1.3273	0.6882	0.7405	0.8367	0.9849	1.2108
DGR	0.7268	0.8240	0.9315	1.1159	1.3395	0.6882	0.7391	0.8355	0.9852	1.2138
CFM	0.7318	0.7832	0.8832	1.0268	1.1539	0.6826	0.7283	0.8166	0.9323	1.1227

component estimates. The collapsed factor forecasts are then similar to the factor-augmented VAR forecasts of SW. We also find that the relative performance of CFM increases when the forecasting horizon gets higher while the SW forecasts appear to perform worse compared to the AR forecasts. This finding is in line with results where Kalman filter forecasts outperform other forecasts under correct model specification, especially for longer forecast horizons; compare the findings of Marcellino, Stock, and Watson (2006).

Table 2 presents the results for a different DGP that resembles an approximate factor model with heteroskedastic and serially correlated idiosyncratic errors. In this case the principal component estimates are not efficient while Doz, Giannone, and Reichlin (2011) claim that their second step Kalman filter and smoothing estimates may be more accurate. We find that our method improves now stronger compared to the other two factor forecasts but differences remain small with gains of around 5% smaller MSEs. Here, different forecasting horizons also give different results. The short-term forecasts from the collapsed dynamic factor model are similar (sometimes slightly worse) to the SW forecasts. However for larger forecast horizons, the CFM forecasts are better with MSE reductions of around 10%. Moreover, these results become more pronounced for a larger time series dimension T.

We also provide the simulation results for parameter settings that represent a weak factors situation. In this case, we set $c_i = 0.9$ such that the common factors have relatively little

explanatory power. The results are presented in Table 3 and we learn that the SW forecasts have large MSEs compared to the AR forecasts for small sample sizes. On the other hand, the CFM forecasts appear to produce forecasts that have a similar or even slightly better precision than the AR forecasts. It is interesting to find that the mean of the signal-to-noise ratio of the principal component estimates increases to approximately 0.20 for $T = 50$ and $N = 10$. For a larger N and T, and for short-term forecasts, the CFM forecasts are more precise than the AR forecasts even when the factors are weak. In almost all cases the CFM forecasts have the smallest MSE among all other models.

Table 3: Monte Carlo study for model with weak factors
This table presents the MSEs of 5000 h-step ahead forecasts for $h = 1, 2, 3, 6, 12$ and different sample sizes N and T. Parameter values of DGP model reflect factor model with weak factors ($\rho = 0.9$, $d = 0.5$, $\tau = 0.5$, $c_i = 0.9$). Different forecast models are random walk (RW), autoregression of order 2 (AR), the principal component SW method, the smoothed factor DGR method, and our collapsed dynamic factor model (CFM). The smallest MSE for each experiment is highlighted.

	T = 50					T = 100				
	h = 1	h = 2	h = 3	h = 6	h = 12	h = 1	h = 2	h = 3	h = 6	h = 12
			N = 10					N = 10		
RW	1.2436	1.3547	1.4860	1.7309	2.0566	1.1433	1.2477	1.3439	1.5718	1.8749
AR	0.8382	0.8971	0.9434	1.0026	1.1698	0.7762	0.8377	0.8933	0.9955	1.1241
SW	0.8570	0.9391	0.9791	1.0731	1.2727	0.7762	0.8378	0.9071	1.0303	1.1720
DGR	0.8593	0.9379	0.9774	1.0679	1.2661	0.7758	0.8398	0.9072	1.0285	1.1701
CFM	0.8350	0.8902	0.9452	1.0047	1.1626	0.7419	0.8160	0.8791	0.9959	1.1167
			N = 50					N = 50		
RW	1.2169	1.3389	1.4941	1.7448	2.1198	1.1293	1.2710	1.3553	1.5888	1.8761
AR	0.8139	0.8728	0.9294	1.0360	1.2021	0.7759	0.8474	0.8807	1.0136	1.1304
SW	0.8137	0.8887	0.9565	1.0967	1.3197	0.7532	0.8275	0.8833	1.0450	1.1843
DGR	0.8096	0.8850	0.9523	1.0960	1.3065	0.7526	0.8288	0.8810	1.0413	1.1801
CFM	0.8037	0.8554	0.9225	1.0291	1.1983	0.7379	0.8116	0.8579	1.0058	1.1114
			N = 100					N = 100		
RW	1.2780	1.4299	1.4961	1.7997	2.0076	1.1821	1.2842	1.4020	1.6209	1.8806
AR	0.8473	0.9017	0.9488	1.0674	1.1549	0.8037	0.8601	0.9049	1.0292	1.1019
SW	0.8138	0.8889	0.9498	1.1244	1.2621	0.7362	0.8146	0.8827	1.0371	1.1510
DGR	0.8125	0.8885	0.9454	1.1192	1.2547	0.7361	0.8139	0.8813	1.0379	1.1463
CFM	0.8122	0.8734	0.9210	1.0488	1.1482	0.7254	0.8045	0.8669	1.0019	1.0792

5 Two empirical illustrations

In this section we illustrate the forecasting performance of our collapsed dynamic factor analysis relative to a selection of benchmark models for pseudo real-time forecasting of macroeconomic variables in the US economy. We adopt the dataset of Stock and Watson (2005) to forecast monthly industrial production and quarterly GDP based on a panel of 132 indicator variables observed at a monthly frequency from 1960 to 2003. For a detailed data and transformation discussion we refer to the Appendix of Stock and Watson (2005)). We compare the forecast accuracy based on the MSE and the mean absolute error (MAE)

using recursive parameter and factor estimates. We do not consider issues with respect to publication delays and/or data vintages. We focus on the final revised data sets in our forecasting exercise.

5.1 Forecasting Monthly Industrial Production

We first consider the forecasting performance of our collapsed dynamic factor model (CFM) for target variable US industrial production. The macroeconomic panel and target variables are all observed at a monthly frequency. Table 4 presents the MSE and MAE of the h-step ahead forecasts produced by the collapsed dynamic factor model and compares it with those of the four benchmarks RW, AR, SW and DGR. For the AR forecasts we adopted the autoregressive model of lag order 2. The same abbreviations for the four methods are used as in the previous section.

Dynamic factor models are typically used for short-term forecasting and therefore we focus on the forecast horizons $h = 1, \ldots, 6$. We also consider a subsample of the entire dataset starting with the first forecasts in 1990 so that T is somewhat small. For the large and small time dimensions, the estimation window starts 10 years earlier. For the factor models, we chose the number of factors according to the criterion proposed by Bai and Ng (2002). As reported in previous studies which use the same data set, the Bai and Ng criterion obtains its minimum for 7 factors; see also the findings in Stock and Watson (2005) and Jungbacker and Koopman (2008). However, this number is quite large and for forecasting purposes it may be interesting to consider a smaller number of factors. We therefore also report the results for one, two and three factors.

Table 4 presents the results of our forecasting exercise. The collapsed factor model compares well when compared with SW and DGR; in 11 out of 16 cases it produces the smallest MAE or MSE. This holds for the analysis based on both short and large samples although the gains are more pronounced in small samples. We also evaluated the forecasting performance of CFM versus SW and DGR based on Diebold and Mariano (1995) tests. The cases where CFM produces more accurate forecasts than SW and DGR, at the 15% level, are marked by a star and dagger, respectively. The result shows that improvements are statistically significant in some but only a few cases. Moreover, we find that for the long sample, the MSE are typically higher than MAE; it indicates a presence of large forecast errors (outliers). The one-step-ahead forecasts of CFM perform relatively worse in terms of MSE in many cases while differences in MAE are smaller. The noise-to-signal ratio is close to zero for all factors indicating that principal component estimates are quite accurate for the given dataset and for the selected dimensions of T and N. We also learn that dynamic factor model forecasts based on information from many predictors do not dramatically outperform simple AR forecast; in particular for the longer forecasting horizons. This is an indication that much of the information is contained in the forecasts series itself; similar findings are

Table 4: Forecast comparisons for monthly US Industrial Production

We report MSE and MAE statistics for different forecast models including random walk (RW), autoregression of order 2 (AR), the principal component SW method, the smoothed factor DGR method, and our collapsed dynamic factor model (CFM), and for different number of factors. A short sample with 138 forecasts (January 1990 – June 2003) and a long sample with 378 forecasts (January 1970 – June 2003) are considered. The smallest MSE and MAE over all different models is highlighted. More accurate forecasts by CFM compared to SW and DGR (according to the Diebold-Mariano two-sided test, 15% significance level) are marked by $*$ and †, respectively.

	Short sample (1990-2003)								Long sample (1970-2003)							
	h = 1		h = 2		h = 3		h = 6		h = 1		h = 2		h = 3		h = 6	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
RW	1.4143	0.9105	1.0296	0.7829	1.0204	0.7448	0.9918	0.8050	1.3259	0.8746	1.2074	0.8329	1.2147	0.8346	1.4013	0.8872
AR	0.6514	0.6524	0.6083	0.6206	0.6403	0.6392	0.6781	0.6793	0.7538	0.6581	0.7287	0.6410	0.7511	0.6517	0.8100	0.6737
1 factor																
SW	0.6235	0.6329	0.6100	0.6273	0.6362	0.6366	0.6785	0.6749	0.6617	0.6184	0.6922	0.6276	0.7366	0.6440	0.8341	0.6797
DGR	0.6025	0.6202	0.6060	0.6259	0.6340	0.6353	0.6817	0.6763	0.6570	0.6137	0.6982	0.6288	0.7412	0.6451	0.8251	0.6780
CFM	0.6117	0.6212	0.5955	0.6158	0.6196*	0.6279	0.6490*	0.6625	0.7186	0.6353	0.6841	0.6214	0.7001*†	0.6357	0.7871	0.6661
2 factors																
SW	0.6008	0.6174	0.6245	0.6344	0.6607	0.6519	0.6772	0.6770	0.6506	0.6199	0.6803	0.6341	0.7183	0.6558	0.7994	0.6864
DGR	0.5882	0.6120	0.6243	0.6359	0.6652	0.6533	0.6795	0.6774	0.6456	0.6220	0.6855	0.6386	0.7254	0.6567	0.8021	0.6901
CFM	0.6097	0.6111	0.6027	0.6140	0.6360†	0.6328†	0.6814	0.6755	0.6853	0.6258	0.6902	0.6175†	0.7348	0.6448	0.8036	0.6672
3 factors																
SW	0.6147	0.6254	0.6227	0.6308	0.6633	0.6537	0.6800	0.6787	0.6380	0.6134	0.6590	0.6229	0.7006	0.6493	0.7770	0.6704
DGR	0.5976	0.6155	0.6234	0.6321	0.6726	0.6572	0.6864	0.6808	0.6315	0.6130	0.6675	0.6313	0.7109	0.6532	0.7797	0.6739
CFM	0.6437	0.6229	0.6044	0.6132	0.6352*†	0.6311*†	0.6809	0.6753	0.6914	0.6291	0.6992	0.6243	0.7438	0.6430	0.8189	0.6710
Bai and Ng (2002) factors																
SW	0.6281	0.6258	0.6161	0.6274	0.6265	0.6365	0.7023	0.6904	0.6361	0.6102	0.7055	0.6378	0.7145	0.6510	0.7776	0.6769
DGR	0.6260	0.6213	0.6251	0.6285	0.6458	0.6446	0.7198	0.6991	0.6398	0.6151	0.7218	0.6528	0.7357	0.6632	0.7793	0.6768
CFM	0.7554	0.6591	0.6016	0.6066	0.6352	0.6286†	0.6801*†	0.6784†	0.7407	0.6499	0.6884	0.6319	0.7324	0.6495	0.8391	0.6810

discussed in Stock and Watson (2006b).

5.2 Forecasting Quarterly Gross Domestic Product

A key macroeconomic variable that summarizes the state of the economy is Gross Domestic Product (GDP) which is collected at a quarterly frequency and is published with a delay of several weeks. To obtain an estimate of the state of the economy before the first official GDP figures are released, recent approaches have emphasized the importance of GDP nowcasting and forecasting using monthly economic and financial variables; see, for instance, Giannone, Reichlin, and Small (2008), Banbura and Rünstler (2011) and de Winter (2011). Here we investigate the performance of the collapsed dynamic factor model to nowcast and forecast quarterly US GDP using a large time series panel of monthly economic variables.

We compare the performance of the CFM with the naive random walk (RW) model and an unobserved components model based on an AR model for unobserved monthly GDP (flow variable) which is linked with observed quarterly GDP via a cumulator variable; we have discussed the details in section 3.5. We further consider two other approaches of making use of information from monthly panels of macroeconomic variables when forecasting a quarterly variable. The first approach is known as the bridge equation forecasts that takes a weighted average of individual indicator forecasts where the weights are functions of the inverse MSEs from past forecasts; see Angelini, Camba-Méndez, Giannone, Rünstler, and

Reichlin (2008) for further details. The second approach is the mixed frequency version of Doz, Giannone, and Reichlin (2011) that is recently proposed to forecast quarterly GDP from monthly indicators by Banbura and Rünstler (2011). We refer to the two approaches as BE and BR, respectively.

Table 5: Forecast comparisons for quarterly US Gross Domestic Product
We report MSE and MAE statistics based on random walk (RW), autoregressive model with cumulator of order 2 (AR), pooled bridge equation (BE), Banbura and Rünstler (2011, BR), and our collapsed dynamic factor model (CFM) forecasts. A short sample with 43 forecasts (Q3 1990 – Q2 2003) and a long sample with 123 forecasts (Q3 1972 – Q2 2003) are considered. The smallest MSE and MAE over all different models is highlighted. More accurate forecasts by CFM compared to BE and BR (according to the Diebold-Mariano two-sided test, 15% significance level) are marked by $*$ and †, respectively.

	Short sample (1992-2003)								Long sample (1972-2003)							
	h = 1		h = 2		h = 3		h = 6		h = 1		h = 2		h = 3		h = 6	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
RW	0.6326	0.6324	0.6476	0.6469	0.6515	0.6529	0.4601	0.5515	1.3102	0.8484	1.3160	0.8490	1.3119	0.8494	1.4948	0.8454
AR	0.3933	0.4805	0.4012	0.4861	0.4138	0.4975	0.3590	0.4627	0.8832	0.6812	0.8818	0.6784	0.8824	0.6809	0.9119	0.6855
BE	0.3003	0.4240	0.3229	0.4405	0.3575	0.4657	0.3506	0.4589	0.7303	0.6177	0.7650	0.6308	0.8339	0.6614	0.8668	0.6682
1 factor																
BR	0.3329	0.4441	0.3438	0.4523	0.3649	0.4668	0.3612	0.4631	0.5441	0.5244	0.5233	0.5268	0.7094	0.6406	0.9945	0.7017
CFM	0.2652	0.4155	0.2919	0.4423	0.3694	0.4797	0.3525	0.4564	0.6270	0.5409	0.5841	0.5325	0.7454	0.6365	1.0544	0.7093
2 factors																
BR	0.4260	0.5148	0.4249	0.5114	0.4230	0.5131	0.4065	0.4969	0.8596	0.6134	0.7446	0.5992	0.8752	0.6869	0.9948	0.7109
CFM	0.2700*	0.4217*	0.2912*	0.4367	0.3597	0.4708	0.3528	0.4609	0.5493*†	0.5270*†	0.5212*†	0.5260*†	0.6706*†	0.6178*	0.9445	0.6793
3 factors																
BR	0.3456	0.4555	0.3318	0.4411	0.3402	0.4527	0.3499	0.4504	0.6369	0.5551	0.6203	0.5556	0.6845	0.6371	0.7756	0.6394
CFM	0.2692	0.4216	0.2877	0.4398	0.3501	0.4723	0.3377	0.4485	0.5621†	0.5301†	0.5344*†	0.5277†	0.6833†	0.6240	0.9435	0.6817
Bai and Ng (2002) factors																
BR	0.4640	0.5431	0.4598	0.5262	0.4696	0.5298	0.4205	0.5114	0.7604	0.5835	0.7190	0.5937	0.9294	0.6929	0.9866	0.7067
CFM	0.2681*	0.4058*	0.2853*	0.4368	0.3548	0.4635	0.3552	0.4596	0.5194*†	0.5011†	0.4671*†	0.5016*†	0.6498	0.6136	0.9495*	0.6718*

Table 5 presents the MSE and MAE forecast statistics of the collapsed factor model and the four benchmark models. We consider h-step ahead forecasts for $h = 1, 2, 3, 6$ months. When $h = 1$, we are in the second month of a quarter and use information of the monthly panel to forecast (nowcast) quarterly GDP which is released next month. When $h = 2$, we are in the first month of a quarter and forecast quarterly GDP that is released two month ahead. The forecasting performance of the collapsed factor model is in many cases more accurate than the considered benchmarks; especially for more than one factor, the Diebold and Mariano (1995) test reveals superior predictive accuracy compared to the two other forecast methods that use information from many predictors. In most cases we find that the model of Banbura and Rünstler (2011), based on factor estimation methods by Doz, Giannone, and Reichlin (2011), cannot produce forecasts of the same accuracy as CFM. When compared with the bridge equation (BE) forecasts, the CFM forecasts are overall more precise.

For the long sample based statistics, we find that MSEs are typically higher than MAEs indicating the presence of large forecast errors. These outliers in our data set may be due to the two recession periods in the 1970s when forecast errors have been typically larger. We can inspect these outlying quarterly GDP observations in Figure 1 against the nowcasts of

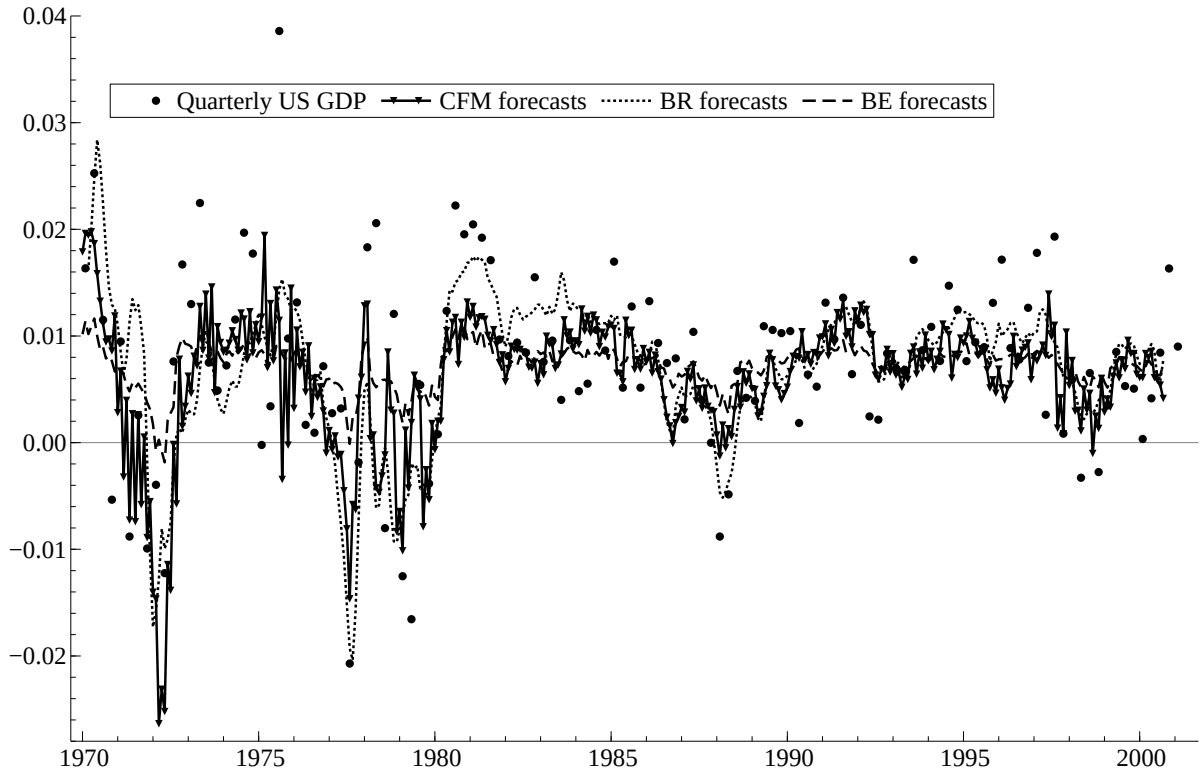

Figure 1: Quarterly US GDP and monthly current quarter nowcasts

Forecasts are from the collapsed dynamic factor model (CFM), the bridge equation (BE), and the Banbura/Rünstler (BR) model forecasts. CFM and BR forecasts are three-month average of monthly latent GDP. The number of factors selection is based on the Bai & Ng criterion.

the current quarter GDP. In the period 1992-2003 we also have many cases where the MAEs are larger than the MSEs. Figure 1 also shows that the forecasts from the three methods are strongly co-moving. However, the two-factor BR and CFM forecasts track quarterly GDP much closer than the pooled bridge equation BE forecasts. It is due to the idiosyncratic factors that the collapsed dynamic factor model forecasts are less smooth than those of Banbura and Rünstler (2011) who only relate quarterly GDP to common factors. Overall we find that the forecasting performance of CFM generally compares well with existing forecasting methods.

6 Conclusions

We have introduced a new method for analysing and forecasting macroeconomic time series when many predictors are available. A fully specified dynamic factor model is collapsed by means of a set of principal components which are simultaneously modeled with the target series. It is shown that the framework of the multivariate unobserved components time series model is instrumental for this development. Most practical issues of macroeconomic forecasting can be handled with our collapsed dynamic factor model. Monte Carlo and empirical evidences are given that show that the collapsed model is a competitive and feasible alternative to the current practices. We leave the development of several extensions of our proposed framework for future research. For example, we can investigate in our framework whether we can improve the analysis and forecasting by having different numbers of principal components and latent factors in the model.

References

Angelini, E., G. Camba-Méndez, D. Giannone, G. Rünstler, and L. Reichlin (2008). Short-term forecasts of euro area GDP growth. Working Paper Series 949, European Central Bank.

Bai, J., E. Ghysels, and J. H. Wright (2011). State space models and MIDAS regressions. Working paper.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*(4), 1133–1150.

Bai, J. and S. Ng (2008). Large dimensional factor analyis. *Foundations and Trends in Econometrics 3*(2), 89–163.

Banbura, M. and G. Rünstler (2011). A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting gdp. *International Journal of Forecasting 27*(2), 333–346.

Breitung, J. and S. Eickmeier (2006). Dynamic factor models. *AStA Advances in Statistical Analysis 90*(1), 27–42.

Chamberlain, G. and M. Rothschild (1983). Arbitrage factor structure, and mean-variance analysis of large asset markets. *Econometrica 51*, 1281–1304.

de Winter, J. (2011). Forecasting GDP growth in times of crisis: private sector forecasts versus statistical models. DNB Working Papers 320, Netherlands Central Bank, Research Department.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–63.

Doz, C., D. Giannone, and L. Reichlin (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics 164*(1), 188–205.

Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

Engle, R. F. and M. W. Watson (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association 76*(376), 774–781.

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*(4), 540–554.

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 M. Lippi (2001). The generalized dynamic factor model: Representation theory. *Econometric Theory 17*(06), 1113–1141.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics 131*, 59–95.

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*(4), 665–676.

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

Harvey, A. C. and A. Jaeger (1993). Detrending, stylised facts and the business cycle. *Journal of Applied Econometrics 8*, 231–47.

Jungbacker, B. and S. J. Koopman (2008). Likelihood-based analysis for dynamic factor models. Tinbergen Institute Discussion Papers 08-007/4, Tinbergen Institute.

Kapetanios, G. and M. Marcellino (2009). A parametric estimation method for dynamic factor models of large dimensions. *Journal of Time Series Analysis 30*(2), 208–238.

Karoui, N. E. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics 36*(6), 2757–2790.

Marcellino, M., J. H. Stock, and M. W. Watson (2003). Macroeconomic forecasting in the euro area: Country specific versus area-wide information. *European Economic Review 47*(1), 1–18.

Marcellino, M., J. H. Stock, and M. W. Watson (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics 135*(1-2), 499–526.

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*(4), 427–443.

Mitchell, J., R. J. Smith, M. R. Weale, S. Wright, and E. L. Salazar (2005). An indicator of monthly GDP and an early estimate of quarterly GDP growth. *Economic Journal 115*(501), 108–129.

Onatski, A. (2009). Asymptotic distribution of the principal components estimator of large factor models when factors are relatively weak. Technical report, Columbia University.

Proietti, T. and F. Moauro (2006). Dynamic factor analysis with non-linear temporal aggregation constraints. *Journal of the Royal Statistical Society, Series C 55*, 281–300.

Stock, J. H. and M. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–79.

Stock, J. H. and M. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statist. 20*, 147–62.

Stock, J. H. and M. Watson (2006a). Forecasting with many predictors. In G. Elliot, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, pp. 515–54. Amsterdam: Elsevier Science Publishers.

Stock, J. H. and M. W. Watson (2005). Implications of dynamic factor models for VAR analysis. NBER Working Papers 11467, National Bureau of Economic Research.

Stock, J. H. and M. W. Watson (2006b). Why has U.S. inflation become harder to forecast? NBER Working Papers 12324, National Bureau of Economic Research.

Stoffer, D. S. and K. D. Wall (1991). Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter. *Journal of the American Statistical Association 86*, 1024–33.