Economic Modelling 29 (2012) 1099–1105

Contents lists available at SciVerse ScienceDirect

Economic Modelling

journal homepage: www.elsevier.com/locate/ecmod

A medium-*N* approach to macroeconomic forecasting

Gianluca Cubadda [a,*], Barbara Guardabascio [b]

[a] *Università di Roma "Tor Vergata", Rome, Italy*
[b] *ISTAT, Rome, Italy*

ARTICLE INFO

Article history:
Accepted 24 March 2012

Keywords:
Partial least squares
Principal component regression
Dynamic factor models
Data-rich forecasting methods
Dimension-reduction techniques

ABSTRACT

This paper considers methods for forecasting macroeconomic time series in a framework where the number of predictors, *N*, is too large to apply traditional regression models but not sufficiently large to resort to statistical inference based on double asymptotics. Our interest is motivated by a body of empirical research suggesting that popular data-rich prediction methods perform best when *N* ranges from 20 to 40. In order to accomplish our goal, we resort to partial least squares and principal component regression to consistently estimate a stable dynamic regression model with many predictors as only the number of observations, *T*, diverges. We show both by simulations and empirical applications that the considered methods, especially partial least squares, compare well to models that are widely used in macroeconomic forecasting.

1. Introduction

Growing attention has recently been devoted to forecasting economic time series in a data rich framework (see, inter alia, Forni et al., 2005; Stock and Watson, 2002a). In principle, the availability of large data sets in macroeconomics provides the opportunity to use many more predictors than those that are conventionally used in typical small-scale time series models. However, exploiting this richer information set comes at the price of estimating a larger number of parameters, thus rendering numerically cumbersome or even impossible the application of traditional multiple regression models.

A standard solution to this problem is imposing a factor structure to the predictors, such that principal component [PC] techniques can be applied to extract a small number of components from a large set of variables. Some key results concerning forecasting with many predictors through the application of PCs are given in Stock and Watson (2002a, 2002b) and Forni et al. (2003, 2005). Recently, Gröen and Kapetanios (2008) have proposed partial least squares [PLS] as alternatives to PCs to extract the common factors. A different methodological framework is Bayesian regression as recently advocated by De Mol et al. (2008) and Banbura et al. (2010). Particularly, these authors attempted to solve the dimensionality problem by shrinking the forecasting model parameters using ridge regression [RR].

A common feature of the mentioned approaches is that statistical inference requires a double asymptotics framework, i.e. both the number of observations *T* and the number of predictors *N* need to diverge to ensure consistency of the estimators. However, an interesting question to be posed is how large the predictor set must be to improve forecasting performances. At the theoretical level, the answer provided by the double asymptotics method is clear-cut: the larger *N*, the smaller is the mean square forecasting error. However, Watson (2003) found that factor models offer no substantial predictive gain from increasing *N* beyond 50, Boivin and Ng (2006) showed that factors extracted from 40 carefully chosen series yield no less satisfactory results than using 147 series, Banbura et al. (2010) found that a vector autoregressive [VAR] model with 20 key macroeconomic indicators forecasts as well as a larger model of 131 variables, and Caggiano et al. (2011) documented that the best forecasts of the 7 largest European GDPs are obtained when factors are extracted from 12 to 22 variables only.

The above results advocate in favor of a sort of "medium-*N*" approach to macroeconomic forecasting. Specifically, we aim at solving prediction problems in macroeconomics where *N* is considerably larger than in typical small-scale forecasting models but not sufficiently large to resort to statistical inference that is based on double asymptotics methods. In order to accomplish this goal, we reconsider some previous results in the PLS literature in a time-series framework. Particularly, we argue that, under the so-called Helland and Almoy condition (Helland, 1990; Helland and Almoy, 1994), both principal component regression [PCR] and the PLS algorithm due to Wold (1985) provide estimates of a stable dynamic regression model that are consistent as *T* only diverges.

Since to date little is known on the statistical properties of PLS in finite samples, a Monte Carlo study is carried out to evaluate the forecasting performances of this method in a medium-*N* environment. To our knowledge, our simulation analysis is unique in that we simulate time series generated by stationary 20-dimensional VAR(2) processes that satisfy the Helland and Almoy condition. Indeed, several studies were devoted to compare PCR and PLS with other methods (see, inter alia, Almoy, 1996) but always in a static framework. Our results suggest that dynamic regression models estimated by PCR and, especially, PLS forecast well when compared to both OLS and RR.

* Corresponding author at: Dipartimento SEFEMEQ, Università di Roma "Tor Vergata", Via Columbia 2, 00133 Roma, Italy. Tel.: +39 06 72595847; fax: +39 06 2040219.
E-mail addresses: gianluca.cubadda@uniroma2.it (G. Cubadda), guardabascio@istat.it (B. Guardabascio).

0264-9993/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.econmod.2012.03.027

In the empirical application, we forecast four US macro time series by a rich variety of methods using similar variables as in the medium dimension VAR model by Banbura et al. (2010). The empirical findings indicate that PLS outperforms the competitors. Interestingly, Lin and Tsay (2006), Gröen and Kapetanios (2008) and Eickmeier and Ng (2011) reached similar conclusions using PLS as an alternative to PCs in large-N dynamic factor models.

The remainder of this paper is organized as follows. The main theoretical features of the suggested methods are detailed in Section 2. The Monte Carlo design and the simulation results are discussed in Section 3. Section 4 compares various forecasting procedures in empirical applications to US economic variables. Finally, Section 5 concludes.

2. Dynamic partial least squares and principal component regression

Let us suppose that the scalar time series to be forecasted, y_t, is generated by the following regression model

$$y_{t+1} = \beta' X_t + \varepsilon_{t+1}, \quad t = 1, \ldots, T, \tag{1}$$

where X_t is N-vector of stationary and ergodic time series, possibly including lags of y_{t+1}, ε_t is a serially uncorrelated error term with $E(\varepsilon_t) = 0$, $E\left(\varepsilon_t^2\right) = \sigma_\varepsilon^2$, $E\left(\varepsilon_t^4\right) < \infty$, and such that $E(\varepsilon_{t+1}|X_t) = 0$. Moreover, we assume that deterministic elements are absent from both time series y_t and X_t, and that each element of X_t has unit variance.

In order to reduce the number of parameters to be estimated in model (1), we follow Helland (1990) and Helland and Almoy (1994) and take the following condition:

Condition 1. *(Helland and Almoy) Let* $\mathbb{E}(X_t y_{t+1}) = \Sigma_{xy}$ *and* $\mathbb{E}(X_t X'_t) = \Sigma_{xx} = \Upsilon \Lambda \Upsilon'$, *where* Υ *is the eigenvector matrix of* Σ_{xx} *and* Λ *is the associated diagonal eigenvalue matrix. We assume that*

$$\Sigma_{xy} = \Upsilon_q \xi, \tag{2}$$

where Υ_q *is a matrix formed by q eigenvectors (not necessarily those associated with the q largest eigenvalues) of* Σ_{xx}, *and* ξ *is a q-vector with all the elements different from zero.*

The above condition is discussed at length in Helland (1990) and Næs and Helland (1993). Essentially, it is equivalent to require that the predictors X_t can be decomposed as

$$X_t = \theta R_t + \theta_\perp E_t,$$

where $R_t = \theta' X_t$, $E_t = \theta'_\perp X_t$, θ and $\theta_\perp$ are, respectively, orthonormal matrices of dimension $N \times q$ and $N \times (N-q)$ such that $\theta\theta' = I_N - \theta_\perp \theta'_\perp$, $E(R_t E'_t) = 0$, and $\Sigma_{xy} = \theta E(R_t y_{t+1})$. R_t and E_t are, respectively, called the relevant and irrelevant components of predictors X_t. The linear combinations $\Upsilon'_q X_t$ that span the space of the relevant components are then called the relevant principal components.

In principle, Condition 1 is in line with the common view that macroeconomic time series are mainly led by few aggregate shocks (e.g. demand and supply shock), which are independent from minor causes of variability (e.g. errors in variables or sector-specific shocks). In Section 4 we will tackle this issue from an empirical viewpoint.

Notice that Condition 1 implies

$$\beta = \Upsilon_q \Lambda_q^{-1} \xi, \tag{3}$$

where Λ_q is the diagonal eigenvalue matrix associated with Υ_q. Hence, model (1) has the following factor structure:

$$y_{t+1} = \xi' F_t + \varepsilon_{t+1},$$

where $F_t = \Lambda_q^{-1} \Upsilon'_q X_t$. Hence, since $\mathbb{E}(y_{t+1}|X_t)$ is a linear transformation of F_t, the predictable component of y_{t+1} is entirely captured by the q components F_t. This is not necessarily the case in dynamic factor models, where the idiosyncratic term is generally not an innovation.[1]

At the population level, PCR computes the prediction for y_{t+1} as $\beta'_{PCR} X_t$ where

$$\beta_{PCR} = \Upsilon_q \Lambda_q^{-1} \Upsilon'_q \Sigma_{xy}. \tag{4}$$

In view of Eq. (3), it is clear under Condition 1 that we have $\beta_{PCR} = \beta$. However, in empirical applications the relevant principal components must be selected and the eigenvalues of the sample covariance matrix of the predictors offer no guidance on this choice. Indeed, Condition 1 does not impose that the eigenvalues associated to the eigenvectors Υ_q are the q largest ones of matrix Σ_{xx} and there is no sound theoretical reason why this should occur (see, inter alia, Hadi and Ling, 1998). As shown by Helland (1990), PLS offer an effective way to overcome this problem.

PLS, introduced by Wold (1985), is an iterative procedure that aims at maximizing the covariance between a target variable and linear combinations of its predictors. In order to accomplish this goal, the first PLS component $\omega'_1 X_t$ is built such that the weights ω_1 are equal to the covariances between the predictors X_t and the target variable y_{t+1}. The second PLS component $\omega'_2 X_t$ is similarly constructed using a new target variable that is obtained by removing the linear effect of the first component on y_{t+1}. In general, the weights of the subsequent PLS factors are set equal to the covariances between X_t and a novel target variable that is obtained by removing the linear effects of all the previously obtained PLS components on y_{t+1}. Hence, let $\beta'_{PLS} X_t$ indicate the prediction of y_{t+1} using the first q PLS components, where

$$\beta_{PLS} \equiv \Omega_q \left(\Omega'_q \Sigma_{xx} \Omega_q \right)^{-1} \Omega'_q \Sigma_{xy}, \tag{5}$$

$\Omega_q = (\omega_1, \ldots, \omega_q)$, and

$$\omega_{i+1} = \Sigma_{xy} - \Sigma_{xx} \Omega_i (\Omega'_i \Sigma_{xx} \Omega_i)^{-1} \Omega'_i \Sigma_{xy}, \quad i = 1, \ldots, N-1 \tag{6}$$

with $\omega_1 = \Sigma_{xy}$. Since it follows by induction from Eq. (6) that Ω_q lies in the space spanned by the eigenvectors Υ_q, it is easy to see that $\omega_i = 0$ for $i = q+1, \ldots, N$ and $\beta_{PLS} = \beta$.

Further features of PLS are better understood by considering the following equivalent way to obtain the weights Ω_q (Helland, 1990). Let us define $V_{0,t} = X_t$ and

$$V_{i,t} = V_{i-1,t} - \phi_i f_{i,t} = X_t - \sum_{j=1}^{i} \phi_j f_{j,t}, \quad i = 1, \ldots, q, \tag{7}$$

where $\omega_i = \mathbb{E}\left(V_{i-1,t} y_{t+1}\right)$, $\phi_i = \mathbb{E}\left(f_{i,t} V_{i-1,t}\right) / \mathbb{E}\left(f'_{i,t} f_{i,t}\right)$, and $f_{i,t} = \omega'_i V_{i-1,t}$ is the i-th PLS factor.

Eq. (7) tells us that the i-th PLS factor $f_{i,t}$ is constructed as a linear combination of the predictors X_t (with weights equal to elements of ω_i) after having removed the linear effects of the previously constructed factors $f_{1,t}, \ldots f_{i-1,t}$. Moreover, by premultiplying each side of Eq. (7) by ω'_i, we see that i-th PLS component can be rewritten as

$$\omega'_i X_t = f_{i,t} + \omega'_i \sum_{j=1}^{i-1} \phi_j f_{j,t}.$$

[1] This property is shared with models obtained through the reduced-rank VAR methodology, see, inter alia, Centoni et al. (2007). However, reduced-rank regression requires the specification of the multivariate model for series (y_t, X'_t) and it is not appropriate for a medium N framework, see inter alia Cubadda and Hecq (2011).