Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 24 (2008) 87 – 100
international journal of forecasting
www.elsevier.com/locate/ijforecast

Macroeconomic forecasting with matched principal components

Christiaan Heij *, Dick van Dijk, Patrick J.F. Groenen

Econometric Institute, Erasmus University Rotterdam, The Netherlands

Abstract

This article proposes an improved method for the construction of principal components in macroeconomic forecasting. The underlying idea is to maximize the amount of variance of the original predictor variables that is retained by the components in order to reduce the variance involved in estimating the forecast model. This is achieved by matching the data window used for constructing the components with the estimation window. Extensive Monte Carlo simulations, using dynamic factor models, clarify the relationship between the achieved reduction in forecast variance and various design parameters, such as the observation length, the number of predictors, and the length of the forecast horizon. The method is also used in an empirical application to forecast eight key US macroeconomic time series over various horizons, where the components are constructed from a large set of predictors. The results show that the proposed modification leads, on average, to more accurate forecasts than previously used principal component regression methods.

Keywords: Time series forecasting; Long term forecasting; Forecast evaluation; Simulation; Dynamic factor model; Factor construction

1. Introduction

One of the basic questions in empirical forecasting is what information should be included in the forecast model. For instance, in many macroeconomic and financial applications, a large number of predictor variables are available. The forecaster then faces the challenge of employing the available information in the best possible way. Various methods for forecasting with many predictors have been proposed in the literature, including forecast combination, model averaging, variable selection, and predictor combination. We refer to Stock and Watson (2006) for a survey. Several empirical studies in macroeconomic forecasting have indicated that, in many cases, the best forecast results are obtained by principal component regression (PCR); see Stock and Watson (1999, 2005) and Banerjee and Marcellino (2006), among others. In PCR, the predictors are summarized by means of a limited number of factor components. As the principal components can be seen as indexes that absorb the main information present in the predictors, one also uses the term ‘diffusion indexes’ for these components.

In this article, we show that further gains in the forecast accuracy of PCR can be achieved by constructing the principal components using a somewhat different method to that usually employed in the

* Corresponding author. Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. Tel.: +31 10 4081269; fax: +31 10 4089162.

E-mail address: heij@few.eur.nl (C. Heij).

0169-2070/$ - see front matter
doi:10.1016/j.ijforecast.2007.08.005

literature. We call our method 'matched PCR' (MPCR), as it matches the two data windows that are used in PCR. More precisely, the distinction between PCR and MPCR lies in the construction of the factors. In PCR, the h-step-ahead forecast made at time T is based on the principal components computed from the (standardized) predictor variables using observations up to and including time T. The forecast model, however, is estimated using the components only up to time T–h. In MPCR, the data window used for constructing the principal components is matched with the data window used for estimating the forecast model by extracting the principal components from the (standardized) predictor variables up to time T–h. This modification better achieves the goal of principal components, namely, to retain the maximum amount of variance of the original predictor variables.

The difference between the two methods, PCR and MPCR, will be most pronounced if the forecast horizon h is large relative to the length of the observation interval T. Put differently, for a fixed horizon h, the distinction between the two methods will vanish if T gets very large. As the two methods are asymptotically equivalent, the merits of MPCR compared to PCR relate to finite sample properties only. The large sample properties of MPCR will be the same as those of PCR, and for those properties we refer to Stock and Watson (2002b) and Bai and Ng (2006).

Although the two methods are asymptotically the same, the differences may be of practical interest in applications where the sample size used for estimation is not so large. In particular, as macroeconomic relationships tend to change over time, PCR is often applied in models that are estimated with a moving window of a restricted length. For instance, if one uses monthly data over a period of ten years to forecast one year ahead, then $T=120$ and $h=12$, so that the intervals employed by PCR and MPCR differ by 10%. The results in this paper show that differences of this magnitude alone may affect the forecast quality considerably.

The article is structured as follows. In Section 2, we outline the current method of forecasting with principal components, and we present and illustrate our method of matched PCR. The relative forecast performances of the original and matched PCR methods are evaluated in Section 3 by means of a simulation experiment, based on Stock and Watson (2002b). Section 4 contains an empirical application involving forecasts of four real economic variables and four price variables, using a set of 146 macroeconomic predictor variables. Section 5 concludes.

2. Forecasting with principal components

2.1. Diffusion index models

In this section, we briefly summarize the method of principal component regression (PCR) proposed by Stock and Watson (1999, 2002a,b), to which we refer for further details. The corresponding forecast models are also called 'diffusion index' models, as the principal components can be interpreted as indexes that summarize the common movements in the underlying macroeconomic predictor variables.

Let y denote the economic variable of interest and let X denote a set of N predictor variables. In PCR, the information in the N predictor variables is summarized by means of k factors f, where k is (much) smaller than N. These factors are used to forecast y by means of a linear regression model. Let h be the forecast horizon and let t denote the current time moment; then the h-step-ahead forecasting model is written as

$$y_{t+h}^{h} = \alpha + \sum_{j=1}^{m} \beta_j' f_{t-j+1} + \sum_{j=1}^{p} \gamma_j y_{t-j+1} + \varepsilon_{t+h}^{h}. \qquad (1)$$

Here y_{t+h}^{h} denotes the h-step-ahead variable to be forecasted. Following Stock and Watson (1999, 2002a, b), we will forecast the h-period average of y_t, so that

$$y_{t+h}^{h} = \frac{1}{h}\sum_{j=1}^{h} y_{t+j}.$$

Model (1) is denoted by DI-AR-Lag. DI-AR is the model without lagged factors ($m=1$), and DI is the model with f_t as the only regressor variable ($m=1$ and $p=0$). If data on y and X is available over a period of length T, then y_{t+h}^{h} can be computed for $t \leq T$–h. The regression in Eq. (1) requires that effective sample size to be at least as large as the number of unknown parameters, so that T–$h \geq 1+km+p$. In particular, this requires that

$$T - h > km. \qquad (2)$$