Economics Letters 125 (2014) 52–56

Contents lists available at ScienceDirect

Economics Letters

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

Robust thresholding for Diffusion Index forecast

Vu Le, Qing Wang *

Department of Mathematics and Statistics, Williams College, Williamstown, MA, United States

HIGHLIGHTS

- There exists serial correlation in the context of long forecasting horizons in Diffusion Index forecast.
- We propose to use robust KVB statistic to replace the regular t statistic in hypothesis tests.
- Hard-thresholding with KVB statistic shows better performance than existing methods.

ARTICLE INFO

Article history:
Received 19 June 2014
Received in revised form
6 August 2014
Accepted 10 August 2014
Available online 19 August 2014

JEL classification:
C1
C5
C12

Keywords:
Diffusion index
KVB
Thresholding
Robust
Serial correlation

ABSTRACT

In this paper we propose a new methodology in improving the Diffusion Index forecasting model (Stock and Watson, 2002a, 2002b) using hard thresholding with robust KVB statistic for regression hypothesis tests (Kiefer et al., 2000). The new method yields promising results in the context of long forecasting horizons and existence of serial correlation. Numerical comparison indicates that the proposed methodology can improve upon the existing hard thresholding methods and outperform the soft thresholding methods (Bai and Ng, 2008) when applied to a real data set that forecasts eight macroeconomic variables in the United States.

1. Introduction

Diffusion Index (DI) forecast has received much attention in recent decades. It offers a fairly simple solution, through principal component analysis (PCA), to combine and extract important information from a large number of variables. This model proves to be increasingly useful in many fields, especially economic forecasting, as large data sets with hundreds of variables have become easily available. From a theoretical perspective, there has been rigorous research, pioneered by Stock and Watson (2002a) and Bai and Ng (2002, 2006), that investigated the asymptotic properties of DI model and proved its consistency. From a practical perspective, practitioners have found substantial values in DI to help improve efficacy of model forecasting. Stock and Watson (2002b) found that DI forecast can largely outperform many models, including autoregressive (AR) model, vector autoregressive (VAR) model, multivariate leading indicator model, and Phillips curve model, when forecasting eight macroeconomic variables in the United States. Using DI model, Brisson and Galbraith (2003) discovered gains in forecast accuracy of low-predictability time series, growth rates of real output and real investment, relative to a variety of alternatives, including the forecasts given by the Organization for Economic Co-operation and Development (OECD).

If we denote y^h_{t+h} as the scalar value of the response variable to be predicted h time units in the future, and let X_t $(N \times 1)$ be the vector of predictor variables, the Diffusion Index model can be expressed as follows:

$$y^h_{t+h} = \alpha + \sum_{j=1}^{m} \beta_j F_{t-j+1} + \sum_{j=1}^{p} \gamma_j y_{t-j+1} + \epsilon_{t+h}, \quad (1)$$

$$X_t = \mathbf{\Lambda} F_t + e_t \quad (1 \le t \le T). \quad (2)$$

Here F_t is a factor of dimension $r \times 1$, β_j is the $1 \times r$ coefficient vector of lag $j-1$ of F_t, γ_j is the scalar coefficient of lag $j-1$ of y_t, ϵ_{t+h} and

* Correspondence to: 18 Hoxsey Street, Bronfman Science Center, Williamstown, MA 01267, United States. Tel.: +1 413 597 4960.
E-mail address: qww1@williams.edu (Q. Wang).

http://dx.doi.org/10.1016/j.econlet.2014.08.010
0165-1765/

e_t are white noises, and y_t and X_t are stationary. In addition, X_t is standardized so that it has zero mean and unit variance. The lags of y_t are included to account for serial correlation. We refer to (1) as the forecast model and (2) as the factor model. The DI forecasting model states that y^h_{t+h} is a linear function of some unobserved factors F_t and its own lags, where F_t can be expressed through a large number of predictors, X_t, with factor loading $\mathbf{\Lambda}$ of dimension $N \times r$.

The factor model (2) can be represented in matrix form as follows:

$$\mathbf{X} = \mathbf{F}\mathbf{\Lambda}' + \mathbf{e},$$

where $\mathbf{X} = (X_1, \ldots, X_T)'$, $\mathbf{F} = (F_1, \ldots, F_T)'$, and $\mathbf{e} = (e_1, \ldots, e_T)'$. Denote $\mathbf{A}$ as the $N \times N$ orthogonal matrix of eigenvectors corresponding to eigenvalues in descending order, and let $\mathbf{A}_r$ be an $N \times r$ matrix with the first r columns of $\mathbf{A}$. The parameters in (2) can be estimated based on Least Squares Criterion by

$$\begin{cases} \hat{\mathbf{\Lambda}} = \sqrt{N}\mathbf{A}_r \\ \hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{\Lambda}}/N. \end{cases}$$

Under identification conditions, Stock and Watson (2002a) show that the principal components consistently estimate the unobserved factors F_t up to a sign. Furthermore, if the usual ordinary least squares (OLS) conditions hold for (1), we obtain asymptotically efficient feasible forecasts (see Theorems 1 and 2 in Stock and Watson, 2002a).

This paper focuses on the practical methodologies in improving the DI forecasting model. Bai and Ng (2008) propose different methods of hard thresholding that can increase forecast accuracy. We noticed the existence of inherently strong serial correlation due to overlapping windows in long forecast horizons, and thus argue against the use of standard t statistic for testing the significance of individual predictor variable in the hard thresholding procedure. We propose to implement the robust KVB test for regression hypothesis (Kiefer et al., 2000) in the context of hard thresholding. Numerical results confirmed that the use of KVB statistic instead of a regular t statistic in hard thresholding yields substantial improvement in forecast efficiency. We found that hard thresholding with KVB statistic is a promising method to identify targeted predictors for DI forecast with long forecast horizons. Moreover, we further argue that compared to soft thresholding (see Bai and Ng, 2008), hard thresholding is more desirable in the context of DI forecast.

The rest of the paper is organized as follows: In Section 2 we propose a new methodology of hard thresholding with robust KVB test and compare it to the existing hard thresholding methods proposed in Bai and Ng (2008). Section 3 presents a real data analysis in comparing the proposed model with various DI models with or without hard thresholding. In Section 4, we argue that hard thresholding using the robust KVB test is a more desirable refinement compared to soft thresholding in DI forecast. We then conclude the paper with some final remarks.

2. Hard thresholding with KVB

One argument against the DI forecast model is that the original set of predictors may not be all relevant in forecasting y^h_{t+h}. This implies that while the leading principal components can explain the majority of the variation among the predictors, they may not be as efficient as possible in forecasting due to the noises from the irrelevant predictors. Bai and Ng (2008) propose hard-thresholding methods that aim to filter out irrelevant predictors from the original data set. They considered a t-test procedure to determine the significance of each potential predictor by augmenting an AR model with each individual predictor. To adjust the threshold for simultaneous significance tests, they considered Bonferroni-type corrections. In this section, we propose a potential refinement for the hard-thresholding method in Bai and Ng (2008) by using the robust KVB statistic to replace the standard t-test statistic in identifying relevant predictors.

2.1. KVB method

When producing monthly forecasts with 12-month forward horizon, there are 11 overlapping months when predicting the response variable. Therefore, there is strong serial correlation (by construction) in the response variable. Because serial correlation tends to inflate the t-statistic, this creates an inevitable issue in hard thresholding whenever the forecasting horizon is longer than the forecast frequency. The longer the forecasting horizon is, the more serious the issue would be. This problem cannot be resolved by simply adding more lags of y_t because (1) the maximum number of lags p can be less than the forecasting horizon h, and (2) allowing too many lags p may seriously affect the finite-sample property of the test statistic (see Cheung and Lai, 1995). One can possibly address the latter issue by fixing the critical values through Monte Carlo simulation. However, this would be computationally intensive, as we need to conduct simulation for each of the hundreds of predictors per forecasting period. Since at the stage of hard thresholding the inference for the individual predictor is of primary interest, we propose to implement the robust regression hypothesis test developed by Kiefer et al. (2000) to adjust for serially correlated errors. The KVB test statistic is known for its robust finite-sample properties. In addition, it fits well into the context of hard thresholding with long forecasting horizons.

The idea behind KVB test is as follows: consider a simple bivariate regression where X_t and β are scalars and X_t is centered. Note that Kiefer et al. (2000) derive the general method for multivariate case. For the purpose of hard thresholding in DI forecast, we only consider the case with a single predictor each time. Denote the OLS estimate for the parameter as $\hat{\beta}$. Define $S_t = \sum_{j=1}^{t} v_j = \sum_{j=1}^{t} X_j \epsilon_j$ and $\sigma_S^2 = \Gamma_0^2 + \sum_{j=1}^{\infty}(\Gamma_j + \Gamma_j')$, where $\Gamma_j = E(v_t v_{t-j})$. Under weak regularity conditions, we have

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \sigma_S^2/\sigma_x^4\right),$$

where σ_x^2 is the variance for the predictor variable. Although one can estimate σ_S^2 using kernel estimation method as suggested by Newey and West (1987), the resulting test statistic is known to have a poor finite-sample distribution and therefore tends to reject the null hypothesis more often than under normality. Kiefer et al. (2000) consider a test statistic that does not require the estimation of σ_S^2. Moreover, it turns out that this statistic has very good finite-sample properties.

The KVB statistic is defined as

$$t_{\text{KVB}} = \frac{\sqrt{T}(\hat{\beta} - \beta_0)}{\sqrt{\hat{C}\hat{\sigma}_x^2}}, \tag{3}$$

where $\hat{C} = T^{-2}\sum_{t=1}^{T}\hat{S}_t^2$ and $\hat{S}_t = \sum_{j=1}^{t} X_j \hat{\epsilon}_j$. The distribution of t_{KVB} is symmetric at 0 with flatter tails than a regular t statistic, as seen in Fig. 1 of Kiefer et al. (2000). Although the asymptotic distribution of t_{KVB} is non-standard, its asymptotic critical values can be simulated easily. For instance, the 97.5th percentile of t_{KVB} is 6.811 (see Table 1 of Kiefer et al., 2000).

Theoretically, using KVB statistic in hard thresholding to identify significant predictors in DI model is robust to the existence of serial correlation. Therefore, we propose to use KVB statistic to replace the regular t statistic in hard thresholding, and expect that this refinement would improve upon the existing methods. To investigate the performance of our proposed method, we will compare forecast efficiency of various models, including our proposal, based on a real data set that is used to forecast eight macroeconomic variables in the United States.